A Statistical Framework for Post-Fabrication Oxide Breakdown Reliability Prediction and Management

Cheng Zhuo, *Member, IEEE,* Dennis Sylvester, *Fellow, IEEE,* and David Blaauw, *Fellow, IEEE*

***Abstract*—Oxide breakdown has become an increasingly pressing reliability issue in modern very large scale integration design with ultrathin oxides. The conventional guard-band methodology assumes uniformly thin oxide thickness, resulting in overly pessimistic reliability estimation that severely degrades system performance. In this paper, we present the use of limited post-fabrication measurements of oxide thicknesses from on-chip sensors to aid in the chip-level oxide breakdown reliability management. A key challenge, which is the focus of this paper, is precisely predicting and managing the reliability condition of each chip with a limited number of measurements and quantifying the tradeoff between reliability margin and system performance. Given the post-fabrication measurements, chip oxide breakdown reliability can be formulated as a conditional distribution that allows one to achieve a significantly more accurate chip lifetime estimation. The estimation is then used to individually tune the supply voltage of each chip for performance maximization while maintaining or improving the reliability. Experimental results show that, by using 25 measurements, the proposed method can achieve an average of 19% performance improvement, and a 27% maximum for a design with up to 50 million devices, with an average operation time of approximately 0.4 s per chip.**

***Index Terms*—Optimization, oxide breakdown, post-fabrication, reliability, variation.**

I. Introduction

Due to aggressive technology scaling, designing a reliable system has become more challenging than ever [1]. The worsening process variation increases susceptibility of the system to various wear-out mechanisms [2]. Among these reliability issues, oxide breakdown (OBD) has emerged as one of the most pressing concerns. As gate oxide thickness is scaled down to the nanometer regime, the stronger electric field across the gate insulator results in faster formation of a conduction path through the dielectric layer, aggravating the risk of destructive breakdown [3]. Even with the change of gate dielectrics nature (high-k dielectrics), the oxide breakdown still remains a major reliability issue for both interfacial and high-k layers in high-k stacks [4].

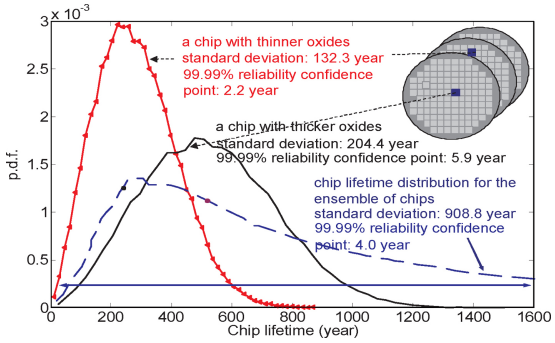


Fig. 1. Chip lifetime distribution for the ensemble of chips (blue dashed curve) with oxide thickness variation of $3\sigma/\mu$ = 4% [8]. The distribution is computed by generating transistor oxide thicknesses that follow the variation model in [10] and then simulating the failure time of each chip in a Monte Carlo fashion.

The conventional worst-case guard-band methodology analyzes chip OBD reliability by assuming a minimum oxide thickness across the chip and then sets a supply voltage level to ensure the required lifetime of the chip. Clearly, such a strategy is overly pessimistic and enforces an overly low supply voltage for the ensemble of chips, causing significant penalty in the performance budget [2], [3]. In practice, no two transistors are identical or have precisely the same characteristics. Instead, they vary significantly from wafer to wafer, reticle to reticle, die to die, and across the die. Thus, some dies with thinner than average oxides are much more likely to fail than others. To more accurately account for the impact of thickness variation on lifetime prediction, there have been recent works incorporating both inter- and intra-die variations into a statistical lifetime analysis [5], [6] or accounting for the circuit functionality and actual stress modes [7].

However, without measurement, designers are not able to know the oxide thickness of an individual transistor on a particular die or determine the specific lifetime expectation from one chip to another. The methods in [5]–[7] or Monte Carlo simulation only relies on the general process variation knowledge, which results in a more accurate but still highly spread lifetime distribution for any chip. This is partly due to the lack of unique information of a particular chip. In other words, it may unfairly imply that a chip with thicker oxides bears the same risk to failure as the one with thinner oxides. Fig. 1 presents the simulated chip lifetime distribution (blue dashed curve) of 50 000 chips. The lifetime spread is

Manuscript received May 31, 2012; revised September 4, 2012; accepted November 9, 2012. Date of current version March 15, 2013. This work was supported in part by the National Science Foundation. This paper was recommended by Associate Editor Y. Cao.

C. Zhuo was with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA. He is now with Intel Corporation, Hillsboro, OR 97124 USA (e-mail: cheng.zhuo@intel.com).

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: dennis@eecs.umich.edu; blaauw@umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2012.2228303

partly from the innate randomness of the OBD mechanism, but further increased by thickness variation. The variation has an exponential effect on the tunneling current and injected charge [8], [9]. This eventually leads to the lognormal shape with a long tail of 908.8-years standard deviation and a 99.99% reliability confidence point of 4.0 years.[1] Without any other information, those numbers will be considered the reliability for all the chips of the design and used to determine the maximum supply voltage. However, each chip has unique oxide thickness conditions for each transistor, and hence some chips are bound to have a significant lifetime margin that could be traded off for higher performance by allowing these chips to operate at a higher supply voltage.

If the oxide thickness of each individual transistor on a fabricated chip can be measured, the lifetime distribution for that chip would be significantly tightened. Fig. 1 also shows the reliability for two particular chips, one with thinner oxides (red curve with triangles, 132.3 years standard deviation/99.99% reliability confidence point is 2.2 years) and one with thicker oxides (black solid curve, 204.4 years standard deviation/99.99% reliability confidence point is 5.9 years). This can be noted from the figure.

1) The estimation based on the blue dashed curve may cause 50%–100% difference from the actual condition.
2) The chip with a thinner oxide (red curve with triangles) has a significantly higher risk to fail early and should be operated with a lower maximum supply voltage, thereby improving the overall reliability of the design. Conversely, the chip with a thicker oxide (black solid curve) is less prone to failure and can be operated at a higher supply voltage limit and hence obtain a performance gain while still meeting the reliability target.

Thus, understanding the oxide thickness condition on a die can result in both performance improvement and higher reliability.

Unfortunately, obtaining the oxide thickness condition for all devices on a die is impossible in today's chips with hundreds of millions to billions of transistors. Recent advances in compact oxide thickness sensors [11], [12] allow tens to hundreds of sensors to be placed on a chip or even inside cores. Thus, the key challenge, which is the focus of this paper, is: how to precisely predict and manage the reliability condition of each chip with a limited number of on-chip oxide thickness measurements. This problem is nontrivial.

1) First, while the number of measurements is limited, the number of transistors on a die in today's technology can be enormous, exceeding 1 billion. Therefore, it is crucial to fully utilize the measurement information to predict the oxide thickness for all devices as accurately as possible.
2) Second, while we can measure the oxide thickness of sensor device with reasonable accuracy, the thicknesses of all other transistors remain uncertain and must be modeled as random variables. Even with a fixed oxide thickness, the reliability for a device itself is a random function representing the probability the device can survive to a certain lifetime [3]. The measurement-driven chip reliability estimation, therefore, turns out to have the form of a conditional multidimensional nested stochastic process. Simple Monte Carlo simulation must model both the random variation in oxide thickness and the innate variation of OBD reliability itself and is therefore extremely expensive in both time and memory.
3) Third, due to the ultrathin oxide thickness in today's processes, the discrete number of atoms along the vertical dimension introduces a random, spatially uncorrelated component to the oxide thickness variation. Hence, oxide thickness across the chip shows noncontinuousness and random nature from transistor to transistor. Recent studies [13] on compressive sensing have exploited the sparsity in frequency domain and achieved deterministic process parameter estimation with a few measurements. However, the independent variation and measurement noise may induce high frequency components (nonsparsity) in the frequency domain and, hence, limit the efficiency of compressive sensing to be applied in OBD reliability.

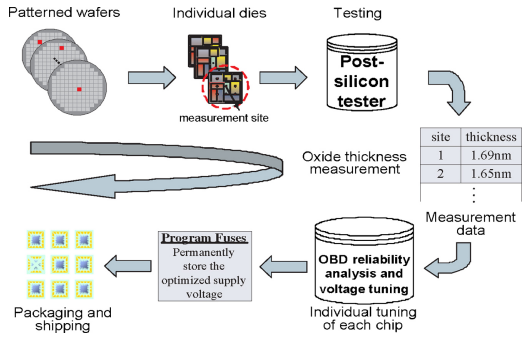


Fig. 2. Proposed post-fabrication oxide thickness measurement-driven supply voltage optimization flow.

In this paper, we propose a new statistical framework for post-fabrication OBD reliability prediction and management using a limited number of measurement points. The measurements of oxide thicknesses for a subset of devices can be conducted by on-chip sensors [11] or test structures [12], which can easily be modified to assess the initial oxide thickness instead of monitoring the degradation process.[2] Fig. 2 illustrates the proposed post-fabrication flow, including the OBD reliability prediction module using the introduced OBD analysis. For each fabricated chip, the measurement is performed once during post-silicon testing to find the initial oxide thickness at the start of its lifetime. Then, the optimal supply voltage limit is selected by the prediction module to maximize performance while maintaining or improving chip OBD reliability. Given the computed supply voltage limit, the tester permanently stores the optimized supply voltage for each chip using either fuses or embedded flash memory. This supply voltage limit is then accessed by the dynamic voltage scaling

[1]The reliability analysis typically requires more than 3σ confidence point due to the large volume of transistors and chips [24]. The 99.99% reliability confidence point is defined as the time when the first 0.01% chips fail.

[2]For example, the sensor in [11] monitors the breakdown leakage, which is exponentially dependent on the oxide thickness. The same sensor can be used to collect the first few cycle data and calibrate the initial oxide thickness. The process-dependent data can be characterized from test chips of the process to enable such calibration.

algorithms and, if available, dynamic reliability management algorithms that control the chip operation during runtime.

The OBD reliability prediction and voltage tuning module in this flow consists of three phases (which are also the major contributions of the proposed work).

1) The first phase uses limited post-fabrication measurements to reduce the uncertainty of the oxide thickness for any unmeasured device. In our framework, we propose to separately account for the inter-die, intra-die spatially correlated, random residual variation components, and the measurement noise. We compute the inter-die component using a maximum-likelihood estimation method and the rest by leveraging the spatial correlation between devices. Then, we construct a conditional distribution based on the post-fabrication measurements, while still preserving the correlation between devices in a conditional covariance matrix.
2) Based on the conditional distribution, the second phase applies conditional principal component analysis to predict the chip reliability [14]. The conditional principal components are employed to derive a tightened lifetime distribution of a particular chip for a given reliability target. The chip lifetime is then bounded by a certain confidence-level interval, the lower bound of which is conservatively used for lifetime evaluation.
3) Finally, in the third phase, we present an optimization flow to efficiently tune the chip maximum supply voltage. As a result, we can boost chip performance for many chips while maintaining or improving reliability.

The remainder of this paper is organized as follows. In Section II, we review the process variation model and oxide breakdown reliability analysis. In Section III, we discuss how to accurately estimate the oxide thicknesses of unmeasured devices by treating inter- and intra-die variation components separately. Section IV details the flow of chip reliability prediction and management for performance maximization, followed by the experimental results in Section V and the conclusion in Section VI.

II. Review of OBD Reliability Analysis

The gate oxide degradation is dependent on oxide thickness, transistor area, supply voltage, and temperature [3]. Although many of the physical details are still under debate, most models note the nondeterministic process of defect generation, eventually resulting in a statistically distributed oxide breakdown time [16], [17] and the strong dependence of this random process on oxide thickness. In this section, we will give a brief review of the oxide thickness variation modeling and previously developed statistical OBD reliability analysis.

A. Oxide Thickness Variation Modeling

The oxide thickness variation can be classified based on the spatial scale over which it manifests [18], [19]. Given the decomposition of global inter-die, intra-die spatially correlated, and random variation components, oxide thickness for any device can be modeled as [10]

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

where u_0 is the nominal oxide thickness. z_g is the inter-die variation component due to the long-range shifts in oxidation temperature and pressure. Clearly, all the devices on the same chip observe the same amount of z_g in oxide thickness, whereas z_g varies for die to die. The fluctuation of z_g among different dies can then be modeled by a Gaussian process $N(0, \sigma_g^2)$ [10]. z_{corr} is the intra-die spatially correlated component that tends to affect closely placed devices in a similar manner. It is typically modeled by a random vector for m devices, $\mathbf{z}_{\text{corr}} = [z_{\text{corr},1}, z_{\text{corr},2}, ..., z_{\text{corr},m}]$, which is a multivariate Gaussian process [10], [18]

$$\mathbf{z}_{\text{corr}} \sim \mathcal{N}_m(0, \Sigma_{\text{corr}}). \tag{2}$$

The subscript of $\mathcal{N}$ denotes the dimensionality of the random vector, and Σ_{corr} is an $m \times m$ covariance matrix for m devices. A simplified spatial correlation model can be achieved by partitioning the chip into N grids and assuming perfect correlation within each grid [18], [19]. Thus, the devices within the same grid have a correlation coefficient of 1 and bear the same spatially correlated variation component, whereas the devices in different grids (i_{th} and j_{th} grids) have a covariance of $\rho_{i,j}\sigma_{\text{corr}}^2$, with a correlation coefficient $\rho_{i,j} < 1$ [18]. The last component z_ϵ in (1) is the independent residual variation resulting from certain local device scale effects, which is modeled as an independent Gaussian process $N(0, \sigma_\epsilon^2)$ [10]. In summary, σ_g, Σ_{corr}, and σ_ϵ denote the uncertainness of the variation components at different spatial scales, which can be either achieved from prior knowledge or extracted from measurements as in [10], [20], and [21].

B. Review of Statistical OBD Reliability Analysis

A common failure criterion for OBD is soft breakdown (SBD), which is initiated by a small gate leakage increase and eventually followed by un-recoverable hard breakdown. SBD is considered an irreversible process with gate leakage increase up to several orders [3]. Some recent works have also noted that the leakage increase does not necessarily lead to circuit or logic failure and circuit may even survive after several breakdowns [3], [7]. Thus, the selection of the failure criterions actually depends on the application or the design under investigation [3]. In this paper, we limit our analysis to determining the initiation of SBD and use this as our failure criteria for large chips, especially CPU designs [3]. In other words, the proposed framework considers the system to have failed as soon as breakdown occurs for any device on the chip.

Due to its stochastic nature, the breakdown time for SBD is modeled as a random variable. The SBD randomness is typically modeled by the Weibull statistics [4], [16], [17], [22]. Even though it is extremely difficult to directly measure the tail cell behavior at a low percentile of the reliability distribution, some earlier works applied Poisson area scaling and found that the tail distribution converges consistently to the Weibull statistics [22], [23]. Thus, in this paper, we follow the use of Weibull statistics to model the innate randomness of SBD [4], [7], [16], [17], [22], [23]

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

where F is the cumulative distribution function (cdf) of time-to-breakdown t, a is the device area normalized with the

minimum device area, and α and β are the scale and shape parameters of the Weibull model. β can be expressed as bx for a given temperature and voltage, where x is the device oxide thickness. The reliability function of a device is then

$$R(t) = P(T > t) = 1 - F(t) = e^{-a(\frac{t}{\alpha})^{bx}}. \quad (4)$$

Since oxide thickness is nondeterministic at the design stage, the device reliability can be interpreted as the conditional reliability given its oxide thickness and can be written as $R(t|x_i)$

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

where x_i is the oxide thickness for the ith device and $f(.)$ denotes the probability density function (pdf). As is discussed in [5] and [6], for a particular chip, if the thicknesses of all devices are known as *a priori* then any device fails independently of all other devices. The overall chip-level reliability is

$$R_c(t) = \int_0^{\infty} \cdots \int_0^{\infty} \prod_{i=1}^{m} R_i(t|x_i) f(x_1 \ldots x_m) dx_1 \ldots dx_m \quad (6)$$

where $\mathbf{x}$ is the vector of oxide thicknesses $(x_1, \ldots, x_m)$, m is the total number of devices on the chip, and $f(x_1, \ldots, x_m)$ is the joint pdf of the gate oxide thicknesses for m devices.

In modern VLSI design, a chip can have millions to billions transistors. To handle the tremendous dimensionality of (6), [5], [6] proposed to project the parametric space of m devices to two distinct random variables, sample mean (u) and sample variance (v) of the chip oxide thickness distribution, which is the frequency distribution of observing certain oxide thicknesses in a particular chip. Based on this, the conditional reliability product $\prod_{i=1}^{m} R_i(t|x_i)$ in (6) with m variables can be simplified to a conditional probability $R_c(t|u, v)$ that only depends on the sample mean u and variance v. The integral of (6) is then compactly expressed as

$$R_c(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_c(t|u, v) f_{\mathbf{uv}}(u, v) du\ dv \quad (7)$$

where $f_{\mathbf{uv}}(u, v)$ is the joint pdf of a Gaussian random variable u and a chi-square random variable v [5]. $R_c(t|u, v)$ for a particular chip can analytically be written as

$$R_c(t|u, v) = \exp[-Ae^{\ln(\frac{t}{\alpha})bu + (\ln(\frac{t}{\alpha}))^2 b^2 v/2}] \quad (8)$$

where A is the die area.

Although the statistical method in [5] and [6] can reduce some pessimism in the traditional guard band [3], the formulation in (7) is still a design time method and only utilizes the general process variation knowledge for a design, without incorporating any post-silicon information. As in Fig. 1, the result obtained by [5] and [6] is still a wide banded distribution. Thus, neither the statistical method [5], [6] nor the guard-band method [3] can distinguish the unique process condition of a particular die. Those methods are still overly pessimistic and result in one global unnecessarily low lifetime estimation for the ensemble of chips.

III. Post-Fabrication Measurement-Driven Oxide Thickness Estimation

In this section, we will present a statistical framework that uses a relatively small number of measurements to significantly reduce the uncertainty of oxide thicknesses for a particular die, and hence provide more accurate lifetime estimation.

A. Problem Formulation

For one particular chip, the inter-die and intra-die variation components (spatially correlated and random) are introduced at different manufacturing stages and hence play very different roles in the oxide thickness model. The inter-die component induces the same increment or decrement to the oxide thicknesses for all the devices within the die and is a constant in (1) for one die. On the other hand, the intra-die spatially correlated and random components vary from device to device. In practice, we cannot distinguish the sources of the variation when the number of measurements is limited. Thus, in analysis, we combine the intra-die variation components together and comprehensively evaluate their impact.

Given a chip dissected to N grids as in [18] with m devices in total, the vector of oxide thicknesses for all the devices is

$$\mathbf{x} = u_0 + z_g + \mathbf{z_{corr}} + \mathbf{z}_\epsilon = u_{\text{chip}} + \mathbf{z_{intra}} \quad (9)$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_m]$ is the oxide thicknesses for m devices, u_{chip}=$u_0 + z_g$ denotes the chip-level oxide thickness mean for this particular chip and may be different from one chip to another, $\mathbf{z}_{\text{corr}}$ is the spatially correlated variation component as in (2), $\mathbf{z}_\epsilon$ is the vector containing the random variation component of each device, $\mathbf{z}_{\text{intra}}$=$\mathbf{z}_{\text{corr}}$+$\mathbf{z}_\epsilon$ is hence the combined intra-die variation component that preserves the spatial correlation among the devices.

Since $\mathbf{z}_\epsilon$ can be interpreted as a multivariate Gaussian process $\mathcal{N}_m(0, \sigma_\epsilon^2 I_m)$, where I_m is an $m \times m$ identity matrix, $\mathbf{z_{intra}}$ is the sum of two multivariate Gaussians and remains a multivariate Gaussian process

$$\mathbf{z_{intra}} \sim \mathcal{N}_m(0, \Sigma_{\text{intra}}) \quad (10)$$

where $\Sigma_{\text{intra}} = \Sigma_{\text{corr}} + \sigma_\epsilon^2 I_m$.

The post-fabrication measurement-driven oxide thickness estimation problem is formulated as follows.

Formulation 1: Given the thickness variation model in (9) and the oxide thickness measurements of n_0 devices across a particular die, estimate the oxide thickness of any unmeasured device (including the components of u_{chip} and $\mathbf{z_{intra}}$) and the corresponding variance.

B. Model Simplification

The grid-based spatial correlation model in [18] indicates that devices within one grid bear approximately the same inter-die and intra-die spatially correlated variation components. This is reasonable when we have relatively finer grids across the chip. The difference in the oxide thicknesses for devices within one grid are then completely attributed to the random variation component, which is independent from one device to another and hence cannot be deterministically predicted. Instead of performing device-level estimation, we employ a grid-based prediction scheme by associating every grid with one random variable. The estimation corresponding to each grid includes: 1) deterministic expected estimation; 2) variance for estimation uncertainty; and 3) the correlation with the estimations for other grids.

Clearly, such modeling simplifies the complexity from the dimensionality of millions (number of devices) to $N + n_0$, where n_0 is the number of sites to be measured and N denotes the number of unmeasured sites with each representing one grid.

After reformulating (9) to the granularity of a grid, both $\mathbf{x}$ and $\mathbf{z}_{\text{intra}}$ are now $(N+n_0)\times 1$ vectors, and

$$\mathbf{z}_{\text{intra}} \sim \mathcal{N}_{N+n_0}(0, \Sigma_{\text{intra,grid}}) \tag{11}$$

where $\Sigma_{\text{intra,grid}}$ is an $(N+n_0)\times(N+n_0)$ covariance matrix for N unmeasured sites corresponding to each grid and n_0 sites to be measured. The entries in $\Sigma_{\text{intra,grid}}$ can be obtained from the covariance matrix Σ_{intra} in (10) by identifying the grids to which the sites belong.

C. Estimating the Chip-Level Oxide Thickness Mean u_{chip}

Before measurements are conducted, the oxide thickness for the sites to be measured remain unknown and hence can be characterized by a multivariate Gaussian model as in (11)

$$\mathcal{N}_{n_0}(u_{\text{chip}}, \Sigma_{mm}). \tag{12}$$

Then, the measurements on n_0 sites $\mathbf{s}=[s_1, s_2 \ldots s_{n_0}]$ can be considered a sample vector drawn from this stochastic model, with measurements acting as n_0 observations. Thus, by using the maximum likelihood estimation (MLE) [24], the maximum likelihood can be achieved when

$$u_{\text{chip}} \approx \frac{[\mathbf{1}]_{1\times n_0}\Sigma_{mm}^{-1}}{[\mathbf{1}]_{1\times n_0}\Sigma_{mm}^{-1}[\mathbf{1}]_{1\times n_0}^{T}}\mathbf{s}^{T} \tag{13}$$

where $[\mathbf{1}]_{1\times n_0}$ denotes a $1\times n_0$ all-one vector. The corresponding MLE estimation variance can be approximately bounded by the Cramér–Rao bound as in [24]

$$\text{var}(u_{\text{chip}}) \approx [\mathbf{1}]_{1\times n_0}\Sigma_{mm}^{-1}[\mathbf{1}]_{1\times n_0}^{T}. \tag{14}$$

D. Estimating the Intra-Chip Variation Component $\mathbf{z}_{\text{intra}}$

If every site of a chip could be measured, the variance for the random vector $\mathbf{x}$ would be reduced to 0. In practice, since the number of measurements is limited, measured oxide thicknesses can only reduce the variance of unmeasured sites to a certain level. In order to assess the impact of measurements, we reorder and separate the oxide thickness vector $\mathbf{x}$ into two subvectors as $\mathbf{x} = [\mathbf{s}, \mathbf{x_u}]$, where $\mathbf{s}$ represents the sites to be measured and $\mathbf{x_u}$ represents the unmeasured sites. $\Sigma_{\text{intra,grid}}$ can then be written as

$$\Sigma_{\text{intra,grid}} = \begin{bmatrix} \Sigma_{mm} & \Sigma_{mu} \\ \Sigma_{um} & \Sigma_{uu} \end{bmatrix} \tag{15}$$

where Σ_{mm} is an $n_0\times n_0$ submatrix containing the covariance for sites to be measured, Σ_{mu} is an $n_0\times N$ submatrix evaluating the covariance between any site to be measured and unmeasured site, Σ_{um} is an $N\times n_0$ submatrix and the transpose of Σ_{mu}, and Σ_{uu} is an $N\times N$ covariance submatrix for unmeasured sites. It is noted that that both subvectors $\mathbf{s}$ and $\mathbf{x_u}$ are multivariate Gaussian variables with a mean of u_{chip} and a covariance matrices of Σ_{mm} and Σ_{uu}, respectively.

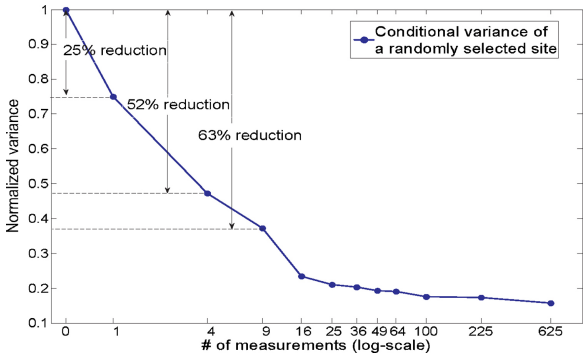


Fig. 3. Reduction in variance of the conditional estimator $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ for a randomly selected site with an increasing number of measurements. The variance is normalized with respect to the variance when no measurement is conducted.

Given the measurement $\mathbf{s} = \mathbf{s_0}$ at n_0 sites, the subvector $\mathbf{x_u}$ for the oxide thicknesses at unmeasured sites can then be expressed in a conditional way, i.e., $\mathbf{x_u}|\mathbf{s} = \mathbf{s_0}$. Such an expression illustrates the impact of measurements on unmeasured sites. By exploiting the spatial correlation between $\mathbf{x_u}$ and $\mathbf{s}$, the pdf for this conditional random vector can be written as

$$f_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}(\mathbf{x_u}) = \frac{f_{\mathbf{x}}(\mathbf{x_u}, \mathbf{s} = \mathbf{s_0})}{f_{\mathbf{s}}(\mathbf{s} = \mathbf{s_0})} \tag{16}$$

where $f_{\mathbf{x}}(\mathbf{x})$ and $f_{\mathbf{s}}(\mathbf{s})$ are pdfs for the multivariate Gaussian random vectors $\mathbf{x}$ and $\mathbf{s}$, respectively, $f_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}(\mathbf{x_u})$ is the conditional pdf for $\mathbf{x_u}$ given $\mathbf{s} = \mathbf{s_0}$. Due to space limitation, we only provide an outline while details can be found in [25], which are also employed in some earlier works [26], [27].

Based on the decomposition of (15), we can define

$$\mathbf{u}_{\mathbf{x_u}|\mathbf{s}} = u_{\text{chip}} + (\mathbf{s} - u_{\text{chip}})\Sigma_{mm}^{-1}\Sigma_{mu} \tag{17}$$

$$\Sigma_{\mathbf{x_u}|\mathbf{s}} = \Sigma_{uu} - \Sigma_{um}\Sigma_{mm}^{-1}\Sigma_{mu}. \tag{18}$$

Thus, given $\mathbf{s} = \mathbf{s_0}$, (16) is still a multivariate Gaussian

$$\mathcal{N}_N(\mathbf{u}_{\mathbf{x_u}|\mathbf{s}}, \Sigma_{\mathbf{x_u}|\mathbf{s}}) \tag{19}$$

where $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ and $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ defined in (17) and (18) are conditional mean and conditional covariance matrix for the conditioned random vector $\mathbf{x_u}|_{\mathbf{s}=\mathbf{s_0}}$.

Intuitively speaking, the vector $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ provides a natural estimation of the oxide thickness at the unmeasured sites, whereas the diagonal entries of $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ evaluate the variance of the estimation. The variance term bounds the remaining uncertainty in the spatial correlation component as well as the independent variation component. According to (18), the conditional variance in $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ is reduced compared with the unconditional variance in (15). Fig. 3 illustrates the trend of variance reduction of the conditional estimator $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ for a randomly selected site from the chip model in Fig. 1 with regards to the growing number of measurements. It is noted that with only nine measurements, the variance of $u_{x_u,i|\mathbf{s}=\mathbf{s_0}}$, as computed in (18), is reduced by 63% compared with the initial variance when no measurement is conducted.

E. Handling Measurement Noise

Electrical measurements can easily be contaminated by measurement noise, which is one of the most fundamental

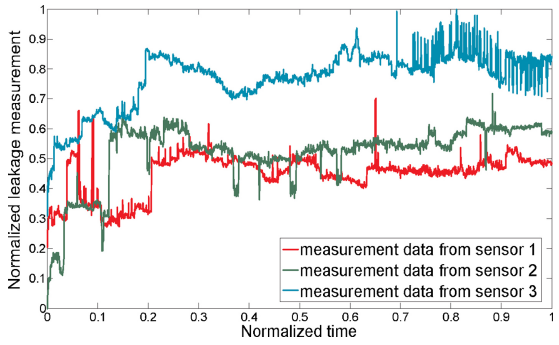


Fig. 4. Gate leakage measurements (normalized) of three OBD sensors in a 45-nm process (the stressed condition is 3.1 V, 100 C).

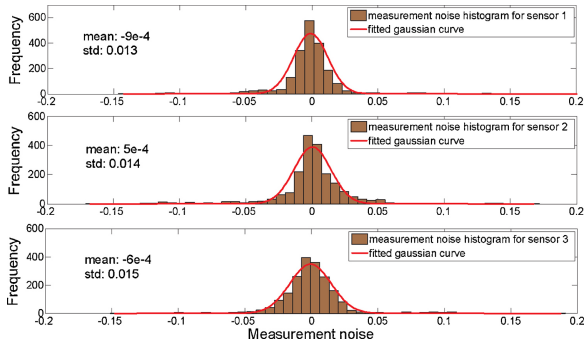


Fig. 5. Histograms and statistics of measurement noises from extracted from three OBD degradation sensors.

problems in post-silicon validation. The presence of measurement errors may bias the data and lead to inconsistent parameter estimation.

In general, measurement noise is treated as a Gaussian distribution according to the central limit theorem. The measurement data from OBD degradation sensors also validated the assumption of Gaussian measurement noise. Fig. 4 shows the measured leakage data of degradation procedure from three OBD degradation sensors [11] at different locations of different test chips.[3] Each data curve represents the measured degradation leakage for a certain period of time and comes with different amount of measurement noise. Since leakage degradation change on the same sensor is continuous with time, the discontinuousness in the curve is highly related to the measurement noise. By denoising the curve using a wavelet filter, the histograms of the extracted measurement noises for all three curves are illustrated in Fig. 5. The extracted measurement noise has consistent close-to-zero mean and around 1% standard deviation.

Based on the calibrated noise model above, assume that the noise on each measurement site is modeled by

$$e \sim \mathcal{N}(0, \sigma_e). \tag{20}$$

The measured data is the superposition of the measurement noise and actual data. This random variable should be added to the oxide thickness variation formulation to account for the noise. The oxide thickness variation model is then

$$\mathbf{x} = u_0 + z_g + \mathbf{z_{corr}} + \mathbf{z}_\epsilon + \mathbf{e} = u_{\text{chip}} + \overline{\mathbf{z_{intra}}} \tag{21}$$

where $\mathbf{e}$ is the measurement noise vector. For the entry corresponding to the measurement site, it follows (20), while for the other entries, it is 0. Similar to (9), we have $\overline{\mathbf{z}_{\text{intra}}} = \mathbf{z}_{\text{corr}} + \mathbf{z}_\epsilon + \mathbf{e}$ to denote the intra-die variation after accounting for the measurement noise, where $\Sigma_{\overline{\text{intra}}} = \Sigma_{\text{corr}} + \sigma_\epsilon^2 + \sigma_e^2$. Since the noise is considered a Gaussian, $\overline{\mathbf{z_{intra}}}$ is still a multivariate Gaussian and the techniques in the earlier subsections can still be applied for oxide thickness characterization.

Procedure: *Chip-Level Oxide Thickness Mean Refinement*
Input: measurements $\mathbf{s}_0$, estimated chip-level oxide thickness mean u_{chip}, the estimator vector $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$, tolerance ϕ
Output: updated chip-level oxide thickness mean u_{chip}

```
Compute sample mean x̄_{N+n0} using (22);
While ||x̄_{N+n0} − u_chip|| ≥ φ
    u_chip = x̄_{N+n0};
    Compute u_{x_u|s=s0} using (17);
    Compute sample mean x̄_{N+n0} using (22);
End while
```

Fig. 6. Algorithm for chip-level oxide thickness mean refinement.

F. Chip-Level Oxide Thickness Mean Refinement

Due to the limited number of measurements, the initial MLE for u_{chip} in (13) may not be the best estimator. By using information from the measurements and the estimator in (17) for the unmeasured sites, the chip level oxide thickness mean u_{chip} can be further refined. In theory, u_{chip} equals the sample mean of all the sites, denoted as $\bar{x}_{N+n_0}$

$$\bar{x}_{N+n_0} = \frac{\mathbf{s_0} \times [\mathbf{1}]_{n_0 \times 1} + \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} \times [\mathbf{1}]_{N \times 1}}{N + n_0}. \tag{22}$$

The deviation between $\bar{x}_{N+n_0}$ and u_{chip} could be attributed to both the estimator error and statistics error. Thus, we can perform a refinement step iteratively to reduce the deviation to a negligible level, i.e., to make $u_{\text{chip}} \approx \bar{x}_{N+n_0}$ by repeatedly replacing u_{chip} in (17) with $\bar{x}_{N+n_0}$ and then computing $\bar{x}_{N+n_0}$ with (22), as shown in Fig. 6. In general, the refinement is completed within tens of iterations to reach certain tolerance, e.g., 10^{-5}. Moreover, it is worthwhile noting that either the estimation variance in (14) or the conditional covariance matrix in (18) does not rely on u_{chip} and remains unchanged for the updated chip-level oxide thickness mean. Thus, the complexity of each iteration is linear with $N + n_0$.

G. Summary of Post-Fabrication Measurement-Driven Estimation

We summarize the algorithm for measurement-driven oxide thickness estimation in Fig. 7. The complexity of the procedure is very low as most computations are analytically achievable. The matrix inverse Σ_{mm}^{-1} and matrix product in (18) are two operations with relatively higher complexly, which depend on the spatial correlation structure of the design and only need to be computed once for a particular design with fixed measurement sites. Since the number of measurements is limited to fewer than hundreds, those operations can numerically be computed within seconds. It is noted that the algorithm results

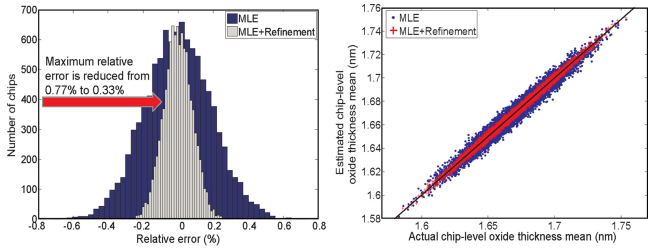
[3]The data was collected for the whole degradation procedure to calibrate the measurement noise. The proposed methodology only requires the initial state data from the sensors (the first few cycles), which takes limited test time.

Procedure: *Post-Fabrication Measurement-Driven Estimation*

Input: measurements $\mathbf{s_0}$, process variation model in (9)
Output: Oxide thickness estimation for each device and the corresponding estimation variance

1: Simplify the model as in subsection III-B to achieve (11);
2: Compute the chip-level oxide thickness mean and corresponding variance using (13) and (14);
3: Estimate the intra-chip variation component $\mathbf{z_{intra}}$ using (17)-(19);
4: Perform chip-level oxide thickness mean refinement as figure 6;
5: Map the estimation and corresponding variance at the granularity of grid level to the devices in the same grid;

Fig. 7. Flow for post-fabrication measurement-driven estimation.



(a) (b)

Fig. 8. Accuracy of chip-level oxide thickness mean estimation. (a) Histograms of relative errors for MLE in Section III-C and MLE with refinement (MLE+Refinement) in Section III-F. (b) Scatter plots for MLE and MLE+Refinement (nominal oxide thickness is 1.67 nm).

in one random variable for each grid, which represents all the devices in the grid. In other words, the estimation for the variable will be projected to all the other devices within the same grid to compute the chip reliability.

We apply the proposed algorithm in Fig. 7 to 10 000 sample chips in 65-nm technology. Each chip has 0.5 million devices, the oxide thicknesses of which are generated by Monte Carlo sampling according to the variation model in [10] and feature parameters in [8]. The chip is then imposed a 50×50 (=2500) grids with 100 uniformly distributed sample sites.[4] The estimated chip-level oxide thickness mean u_{chip} is compared with the actual mean of the oxide thicknesses for all the devices in Fig. 8. From either the histogram or the scatter plot, it can be seen that the estimation achieved by the MLE in Section III-C is very accurate with a maximum relative error of 0.77%, while the mean refinement algorithm (in Section III-F) can further reduce the relative error to a maximum of only 0.33%. We then examine the estimation accuracy at the device level (achieved by step 5 in Fig. 7) for a randomly selected chip. Fig. 9 demonstrates the contour of the difference between the actual oxide thickness and the estimated thickness mapped from the estimator $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ in (17) for all the devices on a chip. With 100 measurements, the accuracy of the oxide thickness estimation for each device is already very high, with average relative error of 0.59% and maximum relative error of 2.8%. Those errors are mainly due to the unpredictable random residual variation but are bounded by the covariance matrix $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ in (18).

[4]The measurement sites are selected in a chessboard manner.

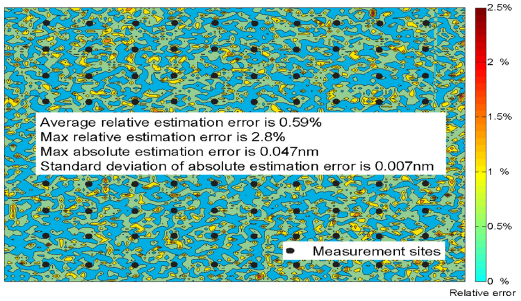


Fig. 9. Contour of the device-level oxide thickness estimation error for a chip with 0.5M devices and 100 samples.

TABLE I
NOTATIONS USED IN OBD RELIABILITY ANALYSIS

Notation	Definition
m	Number of devices in a chip
N	Number of grids in the spatial correlation model
$\mathbf{x} = [x_1, \ldots, x_m]$	The oxide thicknesses for m device of a chip
$\mathbf{x_u}\|\mathbf{s_0}$	The conditional random vector for oxide thicknesses of unmeasured sites, given the measured oxide thicknesses of $\mathbf{s_0}$
$\mathbf{u}_{\mathbf{x_u}\|\mathbf{s}=\mathbf{s_0}}$	Mean of $\mathbf{x_u}\|\mathbf{s_0}$, given measurements $\mathbf{s_0}$
$\Sigma_{\mathbf{x_u}\|\mathbf{s}=\mathbf{s_0}}$	Variance of $\mathbf{x_u}\|\mathbf{s_0}$, given measurements $\mathbf{s_0}$
$\bar{x}_m = \frac{\sum_{i=1}^{m} x_i}{m}$	The sample mean for m devices of a chip
$v = \frac{\sum_{i=1}^{m}(x_i - \bar{x}_m)^2}{m-1}$	The sample variance for m devices of a chip
$R(t_0)$	Chip reliability at time t_0, which is $\Pr(t > t_0)$
T_{target}	Chip design lifetime target
R_t	Chip reliability target at the end of lifetime
T_q	Quantile-based time-to-failure (QTTF)[5], defined as $T_q = \arg_{T_q}\{R(T_q) = \Pr(t > T_q) = R_t\}$
$D_0 = [d_1, \ldots d_N]$	d_i denotes the number of unmeasured devices in the i_{th} grid
$D = diag(D_0)$	A diagonal matrix with diagonal vector of D_0

IV. MEASUREMENT-DRIVEN OBD RELIABILITY PREDICTION AND MANAGEMENT

Based on the oxide thickness estimation discussed in the previous section, we performed a statistical reliability analysis to tighten the lifetime distribution of a chip. Other than the oxide thickness variation, the reliability is also impacted by other variation sources, such as threshold voltage, temperature, etc. Among the variations source, oxide thickness variation is the major factor that impacts the reliability [16], [17], [22], [31], [32]. Here, we focus on chip-level reliability analysis by incorporating the oxide thickness variation and consider the worst-case operating temperature to ensure a correct operation throughout the entire lifetime. By using thermal sensors or process sensors, as in [28] and [29], we can construct the temperature/process profile of the chip and incorporate the other variation sources by performing analysis at the granularity of functional blocks or subblocks, where devices within a block are assumed to bear the same parameter.

For a chip with m devices and N grids for a spatial correlation model, we define the following notations in Table I for the remainder of this paper.

A. Measurement-Driven Reliability Prediction

The challenge of chip-level statistical OBD reliability analysis is the large dimensionality of the integral in (6). Reference [5] proposed to map millions of random variables to two random variables, sample mean and variance of the chip oxide thickness distribution. However, for a conditioned random vector $\mathbf{x_u}|\mathbf{s_0}$, the variables do not bear the same mean and cannot employ the method in [5]. Thus, it is essential to revisit the statistical OBD reliability analysis when measurements are available.

1) *Spatial Correlation Characterization Using Conditional Principal Component Analysis:* Given the measurements $\mathbf{s_0}$, $\mathbf{x_u}|\mathbf{s_0}$ is still a multivariate Gaussian random vector. As in (18), its covariance is

$$\Sigma_{\mathbf{x_u}|\mathbf{s}} = \Sigma_{uu} - \Sigma_{um}\Sigma_{mm}^{-1}\Sigma_{mu}.$$

According to the principal component analysis (PCA), this multivariate Gaussian can be mapped to another set of mutually independent random variables with zero mean and unit variance [5], [18]. Then, for a device in the ith grid, its conditional oxide thickness $x_{u,i}|\mathbf{s_0}$ can be canonically expressed as a linear combination of the principal components

$$x_{u,i}|\mathbf{s_0} = u_{x_u,i|\mathbf{s}=\mathbf{s_0}} + \sum_{j=1}^{N} \lambda_{i,j} z_j \tag{23}$$

where N is the number of principal components (the same as the number of grids in the spatial correlation model), z_js represent the N independent random variables to characterize the spatial correlation, the coefficients $\lambda_{i,j}$s represent the sensitivity of thickness variation with respect to the jth principal component for the random variable in the ith grid. Thus, the conditional random vector of N unmeasured sites can be written compactly with principal components

$$\mathbf{x_u}|\mathbf{s_0} = \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} + \mathbf{z} \times P_\lambda \tag{24}$$

where P_λ is an $N \times N$ matrix containing the sensitivity coefficients $\lambda_{i,j}$s for different principal components and can be achieved by eigenvalue decomposition, $\mathbf{z} = [z_1, z_2, ...z_N]$ is a vector of principal components.

As defined in Table I, the conditional sample mean and sample variance ($\overline{x}_m$ and v) can be expressed in terms of principal components

$$\overline{x}_m = [(\mathbf{x_u}|\mathbf{s_0})D_0^T + \mathbf{s_0} \times [\mathbf{1}]_{n_0\times 1}]/m \tag{25}$$

$$v = \frac{(\mathbf{x_u}|\mathbf{s_0} - \overline{x}_m)D(\mathbf{x_u}|\mathbf{s_0} - \overline{x}_m)^T + (\mathbf{s_0} - \overline{x}_m)(\mathbf{s_0} - \overline{x}_m)^T}{m-1} \tag{26}$$

$\overline{x}_m$ and v describe the characteristics of the conditional chip oxide thickness distribution given measurements $\mathbf{s_0}$. Based on (22), (25) can further be simplified to

$$\overline{x}_m = u_{\text{chip}} + \frac{1}{m}\mathbf{z} \times P_\lambda D_0^T = u_{\text{chip}} + \mathbf{u_{coeff}}\mathbf{z^T} \tag{27}$$

[5]Actual time to failure is a stochastic process and cannot be known until the chip fails. Thus, we introduce a quantile-based time-to-failure which can be interpreted as certain quantile of the time-to-failure distribution. In other words, it is the actual time when chip meets certain reliability target. Note that this value is a deterministic value if the oxide thicknesses of all the devices are known.

where $\mathbf{u_{coeff}} = \frac{1}{m}D_0 P_\lambda^T$. Thus, $\overline{x}_m$ remains a Gaussian

$$\overline{x}_m \sim \mathcal{N}(u_{\text{chip}}, \text{var}(\overline{x}_m)) \tag{28}$$

where $\text{var}(\overline{x}_m) = \text{var}(u_{\text{chip}}) + \mathbf{u_{coeff}}\mathbf{u_{coeff}}^T$.

After expanding the numerator of (25), the conditional variance v can be written as

$$v = \frac{V_{\text{const}} + 2V_1 + V_2}{m-1} \tag{29}$$

where V_{const} is a constant, and V_1 and V_2 are random variables. The formulations can be found in

$$V_{\text{const}} = (\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} - u_{\text{chip}})D(\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} - u_{\text{chip}})^T + (\mathbf{s_0} - u_{\text{chip}})(\mathbf{s_0} - u_{\text{chip}})^T \tag{30}$$

$$V_1 = \mathbf{v_{coeff}}\mathbf{z}^T, \quad V_2 = \mathbf{z}V\mathbf{z}^T \tag{31}$$

$$\mathbf{v_{coeff}} = \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} D P_\lambda^T - (m \cdot u_{\text{chip}} - \mathbf{s_0}[\mathbf{1}]_{n_0\times 1})\mathbf{u_{coeff}} \tag{32}$$

$$V = (P_\lambda^T + [\mathbf{1}]_{N\times 1}\mathbf{u_{coeff}})^T D(P_\lambda^T - [\mathbf{1}]_{N\times 1}\mathbf{u_{coeff}}). \tag{33}$$

Based on (30)–(33), it is found that V_1 is a Gaussian and V_2 has the form of a quadratic normal product, in which V is a positive and symmetric matrix. The uncorrelation between V_1 and V_2, $\overline{x}_m$ and v can be proved in the following two lemmas (the details are presented in the Appendix).

Lemma 1: V_1 and V_2 in (31) are uncorrelated.

Lemma 2: $\overline{x}_m$ in (25) and v in (26) are uncorrelated.

Based on Lemma 1, the mean and variance of v can then be computed from (29) and (30)

$$E(v) = [V_{\text{const}} + tr(V)]/(m-1) \tag{34}$$

$$\text{var}(v) = \frac{2tr(V^2)}{(m-1)^2} + \frac{4}{(m-1)^2}\mathbf{v_{coeff}}\mathbf{v_{coeff}}^T. \tag{35}$$

As detailed in [30], a quadratic normal product random variable as V_2 can be accurately approximated by a chi-square distribution, i.e., $V_2 \sim \hat{a}\chi^2_{\hat{b}}$, with $\hat{a} = \frac{tr(V^2)}{tr(V)}$ and $\hat{b} = \frac{[tr(V)]^2}{tr(V^2)}$, where $tr[.]$ is the sum of diagonal entries. Since the conditional covariance entries in $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ are reduced as in (18), the principal coefficients resulting from the decomposition of $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ are then small and eventually result in a matrix V with small diagonal entries and almost negligible off-diagonal entries. Thus, the degree of freedom for the chi-square distribution $\hat{b} = \frac{[tr(V)]^2}{tr(V^2)}$ is close to N. Given the central limit theorem, the chi-square distribution with a large degree of freedom can be well approximated by a Gaussian distribution [24], [30]. Fig. 10(a) shows a histogram of V_2, with a degree of freedom of 2209 (close to $N = 2500$), the samples of which are collected in a Monte Carlo simulation using (30). The histogram clearly shows Gaussian-like curves with fitting goodness of 0.98 (R-square) and validates the proposed Gaussian approximation. Since both V_1 and V_2 are Gaussians, the un-correlation in Lemma 1 indicates the independence of V_1 and V_2. Thus, v in (29) can also be characterized as a Gaussian random variable.

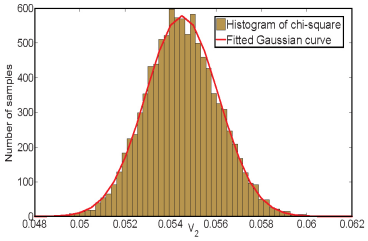


Fig. 10. Comparison of the histogram of chi-square random variable V_2 in (30) with degree of freedom of 2209 (N = 2500) and the fitted Gaussian curve. The fitting goodness is 0.98 (R-square).

2) *Post-Fabrication Measurement-Driven Lifetime Prediction:* Once the underlying distributions of $\overline{x}_m$ and v are characterized, we can conduct the post-fabrication measurement-driven reliability prediction for a particular chip and analyze the quantile-based time-to-failure[6] (QTTF) for a given reliability target R_t by using (8) from [5]

$$R(T_q|\overline{x}_m, v) = \exp[-Ae^{\ln(\frac{T_q}{\alpha})b\overline{x}_m+(\ln(\frac{T_q}{\alpha})b)^2v/2}] = R_t \tag{36}$$

where A is the chip area, α and b are the parameters for the device Weibull reliability function [16], [31]–[33]. This equality illustrates the actual quantile-based time-to-failure when chip meets ceratin reliability target. The QTTF is then compared with design lifetime T_{target} to evaluate chip reliability. To simplify the analysis, we introduce a supplementary random variable $\gamma = \ln(T_q/\alpha)b$ to derive the distribution of T_q (QTTF), and rewrite the equation above as

$$v/2 \times \gamma^2 + \overline{x}_m \times \gamma - \ln(-\ln(R_t)/A) = 0. \tag{37}$$

The solution to (37) is

$$\gamma = \gamma(\overline{x}_m, v) = \frac{-\overline{x}_m + \sqrt{\overline{x}_m^2 + 2\ln(-\ln(R_t)/A) \times v}}{v}. \tag{38}$$

In other words, for a given reliability target R_t, γ is a random function depending on the distributions of $\overline{x}_m$ and v.

By noting that both $\overline{x}_m$ and v have limited variance, we can further simplify (38) by the delta method for first-order approximation [24]

$$\gamma \approx \gamma(E(\overline{x}_m), E(v)) + [\frac{\partial\gamma}{\partial\overline{x}_m}, \frac{\partial\gamma}{\partial v}]|_{E(\overline{x}_m),E(v)} \times [\overline{x}_m - E(\overline{x}_m), v - E(v)]^T. \tag{39}$$

Since both $\overline{x}_m$ and v are Gaussians and uncorrelated, γ follows a Gaussian process with mean and variance

$$E(\gamma) = \frac{-E(\overline{x}_m) + \sqrt{E(\overline{x}_m)^2 + 2\ln(-\frac{\ln(R_t)}{A})E(v)}}{E(v)} \tag{40}$$

$$\text{var}(\gamma) = [(\frac{\partial\gamma}{\partial\overline{x}_m})^2, (\frac{\partial\gamma}{\partial v})^2]|_{E(\overline{x}_m),E(v)}[\text{var}(\overline{x}_m), \text{var}(v)]^T. \tag{41}$$

As $T_q = \alpha\exp[\gamma/b]$, T_q can then be characterized as a lognormal distribution, with mean and variance

$$E(T_q) = \alpha \times \exp(\frac{E(\gamma)}{b} + \frac{\text{var}(\gamma)}{2b^2}) \tag{42}$$

$$\text{var}(T_q) = \alpha^2[\exp(\frac{\text{var}(\gamma)}{b^2}) - 1]\exp(\frac{2E(\gamma)}{b} + \frac{\text{var}(\gamma)}{b^2}). \tag{43}$$

[6] T_q is defined as T_q=arg$_{T_q}${$R(T_q)$=Pr(t>T_q)=R_t} as in Table I. In other words, it is the quantile of reliability distribution for certain reliability target R_t.

B. Reliability Management and Performance Optimization

In practice, the design objective may be a certain design lifetime T_{target} with a predefined reliability requirement R_t, i.e., the probability of chip failures may not exceed $1-R_t$ within T_{target} years lifetime. However, due to process variation, some chips have thinner oxides and are quicker to fail. The tightened QTTF distribution T_q enables us to quantitatively evaluate whether the chip will meet the design lifetime target. The next question is then how much voltage we need to tune to optimize the performance, which will be discussed in the following optimization flow.

Due to the remaining uncertainty of the oxide thicknesses, QTTF itself is a distribution as in (42) and (43). We therefore use the lower bound of the distribution with a certain confidence to ensure a robust design and margin for other variation sources. Conservatively, with a 99.9% confidence level, we can derive the following one-sided confidence interval:

$$T_q \in [\alpha\exp\left[\frac{E(\gamma) - 3\sqrt{\text{var}(\gamma)}}{b}\right], \infty] \tag{44}$$

where the moments of γ can be computed from (40) and (41). The lower bound of (44) is then denoted as T_{lb} and used to evaluate chip lifetime in optimization. In other words, after optimization, we may push the distribution of QTTF to the right of T_{target} and have 99.9% confidence that the chip will meet the lifetime target. Since both parameters α and b in (44) depend on supply voltage [31]–[33], we formulate the following to maximize the supply voltage while T_{lb} meets the design lifetime target.

$$\text{Maximize} \qquad \text{voltage} \tag{45}$$

subject to

$$T_{lb} = \alpha(voltage)\exp\left[\frac{E(\gamma) - 3\sqrt{\text{var}(\gamma)}}{b(voltage)}\right] \geq T_{\text{target}} \tag{46}$$

$$v_{\min} \leq voltage \leq v_{\max} \tag{47}$$

where *voltage* denotes the maximum supply voltage available for the chip, the first constraint in (46) implies that the 99.9% confidence lower bound of QTTF is larger than the design lifetime target, and the second constraint in (47) denotes the possible voltage tuning range. This optimization problem is equivalent to finding the feasible domain of the inequality in (46), where the parameters of the device reliability function ($\alpha(voltage)$ and $b(voltage)$) indicate the underlying dependence on supply voltage. In our implementation, we adopt the linear models as in [31]–[33], and hence achieve a quadratic inequality from (46). As a result, the optimization flow above eventually reduces the failure rate to improve reliability yields, while the performance is also enhanced by reducing lifetime safety margins.

C. Summary of OBD Reliability Prediction and Management

The flow for post-fabrication measurement-driven reliability prediction and management is summarized in Fig. 11. Given n_0 measurements for a particular chip, we first estimate the oxide thicknesses and corresponding variance using a conditional multivariate Gaussian model. The conditional spatial

Procedure: *Post-Fabrication Measurement-Driven OBD Reliability Prediction and Management*

Input: measurements $\mathbf{s_0}$, process variation model in (9), reliability target and design lifetime

Output: optimized supply voltage

1: Given measurements $\mathbf{s_0}$, estimate the conditional oxide thickness and covariance matrix with the flow in figure 7;
2: Apply PCA to the conditional covariance matrix and obtain the underlying distributions of $\overline{x}_m$ and v using (27)-(35);
3: Estimate tightened chip lifetime distribution using (42) and (43);
4: Solve the optimization problem in (45)-(47) to achieve the optimized supply voltage;

Fig. 11. Procedure for post-fabrication measurement-driven OBD reliability prediction and management.

correlation is then explored by conditional PCA to derive the distributions of $\overline{x}_m$ and v, which characterize the underlying conditional chip oxide thickness distribution and help achieve a tightened lifetime distribution. The lifetime estimation then allows an optimization flow to quantify tradeoffs between reliability and supply voltage/performance.

V. Experimental Results

The proposed framework was implemented and tested on several designs using 65-nm devices (nominal oxide thickness is 1.67 nm). The defect generation relationships for the technology node and the technology dependent parameters of the oxide reliability function model are obtained from [16], [31]–[33], which are used in the device-level reliability model. In practice, this can be obtained by a one time per technology characterization using test devices.

Seven designs are employed in the experiment, including six synthetic circuits that were automatically generated and an alpha processor design with 15 functional modules. For each design, we collected 10 000 chips by Monte Carlo simulation that follows the thickness variation model discussed in Section II. According to [8], the $3\sigma/u$ ratio for oxide thickness variation is assumed to be 4% for the nominal value, and then split into 50% global variation, 25% spatially correlated variation, and 25% independent variation [10], [27]. The measurement noise model is based on the calibration of the degradation sensor data in [11] with zero mean and $3\sigma/u = 3\%$. Since the real measurement data for thickness correlation was unavailable, the covariance matrix for thickness variations was derived from an exponential decaying function of the respective distance [10], [20].

A. Efficacy of the Proposed OBD Reliability Prediction

To evaluate the accuracy of the proposed method, the conditional QTTF distribution for a chip was also computed by Monte Carlo simulation with an accept-and-reject strategy. The simulation only accepted sample vectors with similar entries at the sample sites, the tolerance of which was set to 0.01 nm in our implementation. This is equivalent to exploring the parameter space of the conditional random vector $\mathbf{x_u}|\mathbf{s_0}$. The results are shown in Fig. 12 for a design with 0.5 million devices

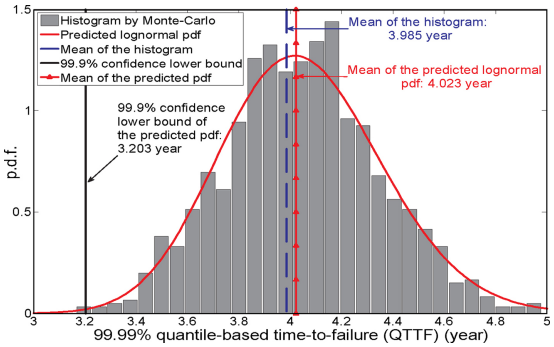


Fig. 12. Accuracy comparison of the quantile-based time-to-failure (QTTF) histogram generated by Monte Carlo simulation and predicted QTTF pdf using proposed method (fitting goodness is 0.96 for R-square). The design has 0.5M devices and 25 samples with 2500 grid cells for spatial correlation modeling. The reliability target R_t is set to 99.99% (100 failures per million).

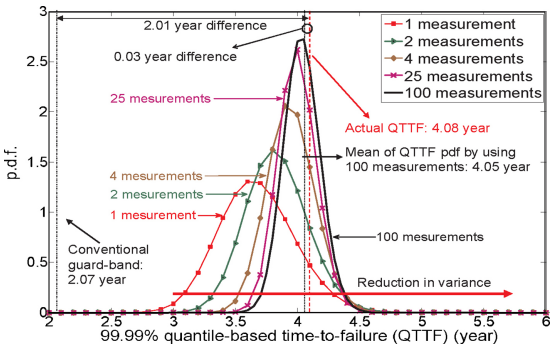


Fig. 13. Reduction in the variance of 99.99% quantile-based time-to-failure (QTTF) distribution for a particular chip with increasing number of samples (1, 2, 4, 25, and 100 samples).

and 25 measurements. That the predicted lognormal pdf using the samples in Sections III and IV shows good agreement with the histogram of Monte-Carlo simulation (1000 sample chips). The difference between the mean of the histogram and lognormal pdf is 0.038 years. The 99.9% confidence lower bound of QTTF is 3.203 years that demonstrates the tightness of the distribution.

For the same chip, we also explored how the predicted QTTF distribution changes when the number of samples is increased. Fig. 13 clearly shows the reduction in variance as the number of samples grows. It is interesting to note that even one or two samples provide sufficient information to tighten the distribution, whereas 100 samples help reduce the standard deviation of the distribution to only 0.16 years. The difference between the actual QTTF point (for a reliability target $R_t = 99.99\%$) and the mean of the predicted QTTF distribution (using 100 samples) is only 0.03 years (0.8% estimation error), while the conventional guard band is 2.07 years with almost 50% estimation error.

Moreover, we studied the convergence of the mean and 99.9% confidence lower bound ($\mu - 3\sigma$) of the predicted QTTF distribution to the exact values, as shown in Fig. 14 for two chips, one with thicker oxides and another with thinner oxides. For each particular sample number, we picked ten different configurations (placement) of sample sites and then computed the 10 set mean/99.9% confidence lower bound of QTTF distribution to achieve the error bar. Note that the exact value can only be obtained by having complete knowledge of

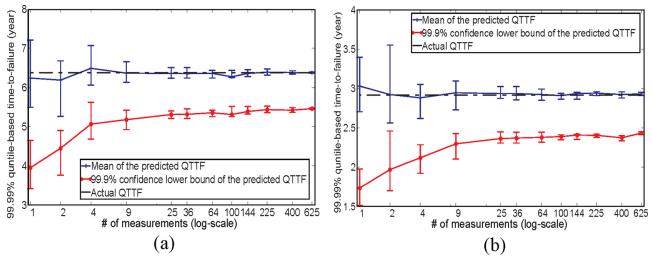


Fig. 14. Convergence of the mean and 99.9% confidence lower bound of the predicted quantile-based time-to-failure (QTTF) distribution with increased samples. (a) Chip with thicker oxides. (b) Chip with thinner oxides.

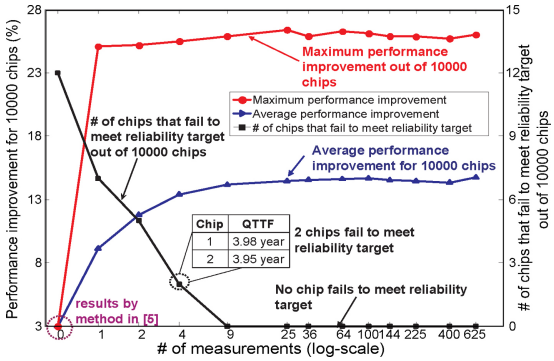


Fig. 15. Reliability management results with increased samples for 10 000 chips of a 0.5M-device design: performance improvement and number of tuned chips that fail to meet target after optimization. Conventional guard-band is employed as baseline for comparison. The results for 0 samples in the figure are from the method in [5]. Others are from the proposed method.

all transistor oxide thicknesses on the chip. When only one or two samples are available, the results may be dominated by the randomness of sample site and have a relatively larger variation. However, with an increasing number of samples, both the estimated values or their variance converge quickly.

B. Reliability Management and Performance Optimization

In this subsection, we applied the proposed post-fabrication measurement-driven methodology to tune the supply voltage of 10 000 chips of a synthetic 0.5M-device design to ensure reliability while maximizing performance. The lifetime target T_q was set to four years for a reliability target R_t = 99.99% and the supply voltage tuning range is 0.8 V to 1.3 V. The relationship between oxide thickness, voltage, and performance is calibrated by SPICE simulation on 65 nm standard cells.

Fig. 15 shows the tuning results using a conventional guard-band design-time statistical analysis in [5] and [6] (denoted as 0 sample in the figure) and the proposed methodology using different numbers of samples. The guard band that assumes minimum oxide thickness across the chip achieved a single supply voltage for all the chips (0.858 V) and was employed as the baseline for comparison. The other two methods used 99.9% confidence lower bound of the predicted QTTF distribution as the evaluation of chip's lifetime. Since [5] uses a more accurate model of the oxide variation compared to the baseline approach, it assigns the ensemble of chip a slightly higher supply voltage of 0.875 V. However, since it is unaware of the unique condition of each particular chip, it remains overly pessimistic and results in a merely 3% performance improvement. On the other hand, with only 25 samples, the

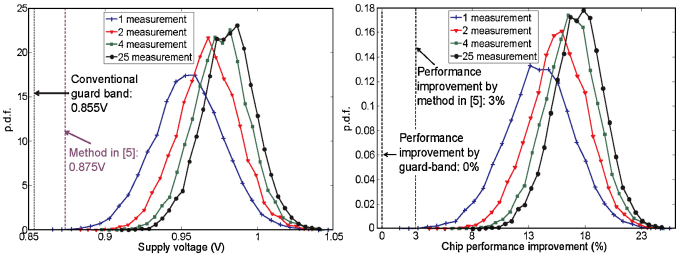


Fig. 16. Distributions of (a) optimized supply voltages and (b) performance improvement with different numbers of samples.

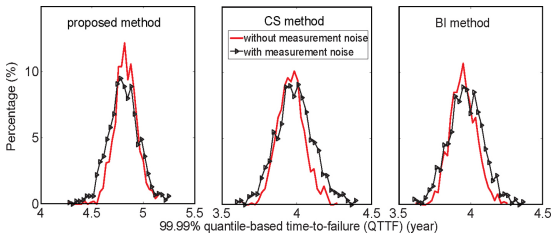


Fig. 17. Comparison of the quantile-based time-to-failure distribution of the chips optimized by the proposed framework, bilinear interpolation based method, and compressive sensing based method.

proposed methodology can obtain a well-tightened QTTF distribution and a more precisely optimized voltage for each chip, achieving 15% performance improvement on average and 26% improvement at maximum.

Moreover, although [5] since the proposed methodology provides more accurate prediction it quickly reduces the number of failures to 0 with increased samples. Even for those chips that fail to ensure reliability, their QTTFs are very close to the design lifetime target. As shown in the figure, with a four-year-target and four samples, two tuned chips fail to meet lifetime target with QTTFs of 3.98 and 3.95 years, respectively. Fig. 16 presents the distributions of optimized supply voltage and the resulting performance improvement using different numbers of samples in tuning. Both plots show a shift to the right with increased samples, indicating the capability to choose a more reasonable supply voltage when more information is available.

C. Reliability Management With Measurement Noise

In this subsection, the proposed framework is applied to the data with measurement noise. The noise model is calibrated based on the degradation data from the degradation sensors [9], [11], which follows a zero mean Gaussian with $3\sigma/u$ = 3%. The estimation accuracy and reliability management results of the framework are compared with two other methods.

1) Bilinear interpolation (BI) based method uses bilinear interpolation to estimate the device oxide thickness.
2) Compressive sensing (CS) based method [13] uses compressive sensing to estimate the device oxide thickness.

Both bilinear interpolation and compressive sensing utilize the smoothness in either the spatial domain or the frequency domain to achieve point estimate, but do not bound the uncertainty of the estimation. Table II summarizes the oxide thickness estimation results by the three methods under two scenarios: sample data with measurement noise and without

TABLE II

OXIDE THICKNESS ESTIMATION ACCURACY COMPARISON OF THE PROPOSED FRAMEWORK, BILINEAR INTERPOLATION (BI) BASED METHOD, AND COMPRESSIVE SENSING (CS) BASED METHOD FOR SCENARIOS WITH AND WITHOUT MEASUREMENT NOISE

Design	#dev.	Without Measurement Noise						With Measurement Noise					
		Average Relative Error			Std. Relative Error			Average Relative Error			Std. Relative Error		
		Proposed	CS	BI	Proposed	CS	BI	Proposed	CS	BI	Proposed	CS	BI
A	80K	0.9%/1×	1.2×	1.3×	0.7%/1×	1.2×	1.3×	2.4%/1×	1.1×	1.1×	1.6% /1×	1.2×	1.3×
B	500K	0.9%/1×	1.1×	1.2×	0.7%/1×	1.1×	1.2×	1.0%/1×	1.2×	1.3×	0.7%/1×	1.2×	1.3×
C	10M	1.0%/1×	1.1×	1.2×	0.7%/1×	1.1×	1.2×	1.0%/1×	1.2×	1.3×	0.8%/1×	1.2×	1.3×
Average		0.9%/1×	1.13×	1.23×	0.7%/1×	1.13×	1.23×	1.5%/1×	1.17×	1.23×	1%/1×	1.2×	1.3×

The results are collected from 10 000 chips with 25 samples per chip. The proposed method is used as baseline for comparison.

TABLE III

FAILURE RATE COMPARISON OF THE PROPOSED FRAMEWORK, BILINEAR INTERPOLATION (BI) BASED METHOD, AND COMPRESSIVE SENSING (CS) BASED METHOD FOR SCENARIOS OF SAMPLES WITH AND WITHOUT MEASUREMENT NOISE

Design	Without Measurement Noise			With Measurement Noise		
	Failure After Optimization			Failure After Optimization		
	Proposed	CS	BI	Proposed	CS	BI
A	0%	59%	63%	0%	65%	69%
B	0%	55%	61%	0%	67%	70%
C	0%	58%	62%	0%	63%	67%
Average	0%	57%	62%	0%	65%	69%

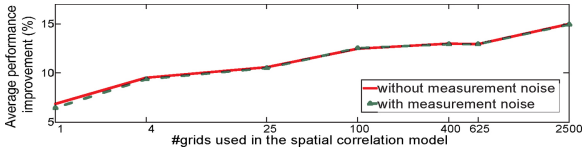


Fig. 18. Impact of the correlation model granularity on the reliability management result for design B (with 25 samples per chip).

measurement noise. Columns 3–8 show the relative estimation error of the three methods for the scenario without measurement noise and columns 9–14 for the scenario with measurement noise. On average, the proposed framework achieves around 0.9% estimation deviation for the scenario without measurement noise and 1% deviation for the scenario with measurement noise. On the other hand, compared with the proposed method, the CS and BI methods show 10%–20% more deviation and 20%–30% more deviation for the two scenarios, respectively.

The thickness estimation is then used to compute the chip reliability and tune the maximum supply voltage by the algorithm in Section IV-B. After tuning, the chip that fails to meet the target reliability for the target life time is considered to be a failure unit. The failure rate of the optimized designs by the three methods are presented in Table III. The proposed method is able to achieve 0 failure across all the designs for the scenarios with and without measurement noise, while the other two methods have 60%–70% failure rate. The large failure rate of other two methods can be attributed to the inaccurate thickness estimation and the lack of scheme to bound the estimation uncertainty. The reliability is exponentially dependent on the oxide thickness. Due to the large number of devices, a small amount of inaccuracy may lead to a much larger reliability prediction error. Without a scheme to bound the estimation uncertainty, the voltage tuning based on such prediction tends to be aggressive and eventually causes

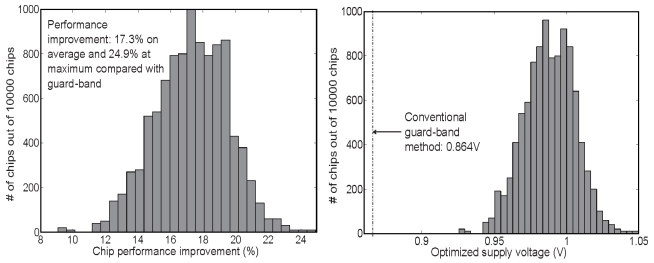


Fig. 19. (a) Performance improvement histogram and (b) optimized supply voltage histogram of 10 000 chips for an alpha-processor-like design with 0.84M devices and 25 samples.

TABLE IV

IMPACT OF MEASUREMENT NOISE ON THE RELIABILITY MANAGEMENT RESULTS BY THE PROPOSED FRAMEWORK

Design	Without Measurement Noise				With Measurement Noise			
	Opt. Volt.		Perf. Impro.		Opt. Volt.		Perf. Impro.	
	Ave.	Std.	Ave.	Max.	Ave.	Std.	Ave.	Max.
A	0.981	0.026	15%	22%	0.979	0.026	14%	22%
B	0.982	0.028	15%	26%	0.980	0.028	15%	26%
C	0.961	0.026	18%	27%	0.960	0.026	18%	27%

significant failure rate. Moreover, the measurement noise may introduce more high frequency components and limit the effectiveness of bilinear interpolation or compressive sensing. Fig. 17 shows the optimized QTTF distributions of design B for the three methods. The QTTF for the scenario with measurement noise is more distributed than the one without noise, as noise increases the uncertainty in the estimation and requires a slightly larger bound to ensure the reliability target.

Table IV summarizes the performance improvement results after reliability management. Columns 2–5 display the average and standard deviation of the optimized voltage, the average and maximum performance improvement (compared with the guard-band method) for the scenario without measurement noise and columns 6–9 display the results for the scenario with measurement noise. For large designs, the impact of measurement noise on the optimized performance is negligible between the two scenarios, as the proposed work is able to accurately identify and bound the uncertainty in the estimation. Fig. 18 compares the impact of the correlation model granularity on the performance optimization result. The solid curve is the case without measurement noise and the dashed curve is with measurement noise. Even with a one-grid correlation model, the proposed method can still achieve around 7% performance improvement on average for both cases.

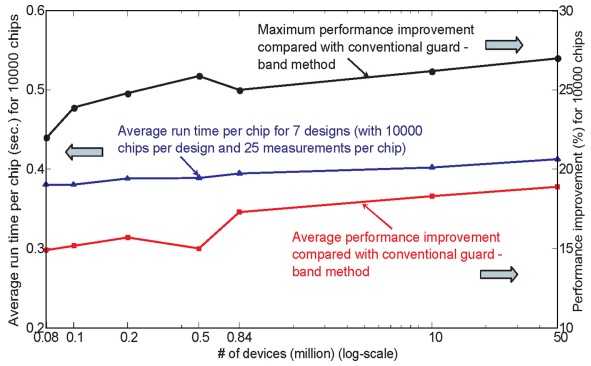


Fig. 20. Average runtime per chip and average performance improvement for seven different-sized designs (10 000 chips for each design and 25 samples per chip). Conventional guard band is employed as a baseline.

D. Scalability

The scalability of the proposed methodology is examined in both its dependence on design complexity/size and runtime. We first applied the approach to an alpha-processor-like design, with 15 functional blocks and 0.84M devices in total. Due to the functional block difference, the grids for spatial correlation model have nonuniform densities, i.e., each grid has different number of devices. We sampled 25 devices per chip and tuned 10 000 chips resulting in a performance improvement of 25% at maximum and 17% on average compared with the guard-band method, as shown in Fig. 19.

We then applied the proposed method to tune 10 000 chips of seven different designs and recorded performance improvement and average runtime per chip. Fig. 20 shows a flat curve of runtime of around 0.38 s, and a slightly growing trend of average performance improvement from 15% to 19% and maximum improvement from 22% to 27%. The methodology runtime relies more on the number of grids for spatial correlation model instead of circuit size as validated in the figure, which is an appealing feature for modern processors with increasingly larger designs.

VI. Conclusion

This paper presented a post-fabrication measurement-driven OBD reliability prediction and management methodology. The methodology used limited measurements to estimate the oxides condition of a chip. The estimation was then incorporated into a statistical model to more accurately predict chip lifetime distribution, which is fed to an optimization flow to trade off reliability margin and system performance. Experimental results showed that even for a design with up to 50 million devices, the methodology can achieve 19% performance improvement on average and 27% at maximum compared with conventional guard-band, while with an average runtime of only 0.4 s.

Appendix

Lemma 1: V_1 and V_2 in (31) are uncorrelated.

Proof: Since z_is are standard independent Gaussians and V_1 is the weighted sum of z_i, $E(V_1) = 0$, where $E(.)$ computes the expected value of a random variable.

For any i, j, k, $E(z_i) = E(z_i z_j z_k) = 0$. Thus, it can be derived from (30) that

$$E(V_1 V_2) = \mathbf{v_{coeff}} E(\mathbf{z}^T \mathbf{z} V \mathbf{z}^T) = 0. \quad (48)$$

As a result, we have $E(V_1)E(V_2)$=$E(V_1 V_2)$, which indicates V_1 and V_2 are uncorrelated. ■

Lemma 2: $\bar{x}_m$ in (25) and v in (26) are uncorrelated.

Proof: Similar to Lemma 1, for any i, j, k, $E(z_i) = E(z_i z_j z_k) = 0$ and $E(z_i^2) = 1$. Thus, it can be derived from (27) and (29) that

$$E(\bar{x}_m)E(v) = u_{\text{chip}} \times [V_{const} + tr(V)]/(m-1). \quad (49)$$

For $E(\bar{x}_m v)$, it is formulated as

$$E(\bar{x}_m v) = \frac{u_{\text{chip}} V_{const} + 2E(\bar{x}_m V_1) + E(\bar{x}_m V2)}{m-1}. \quad (50)$$

By definition, $E(\bar{x}_m V_1)$ is

$$E(\bar{x}_m V_1) = \mathbf{u_{coeff}} \mathbf{v_{coeff}^T}. \quad (51)$$

Since $P_\lambda P_\lambda^T$ is a unit matrix, the equation above can be simplified with (27) and (32)

$$E(\bar{x}_m V_1) = \frac{D_0}{m}(D_0 \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}{}^T + \mathbf{s_0}[\mathbf{1}]_{n_0 \times 1} - m \times u_{\text{chip}}). \quad (52)$$

With the equality of u_{chip} in (22), the equation in the bracket is just 0. Thus, $E(\bar{x}_m V_1) = 0$.

Similar to Lemma 1, $E(\bar{x}_m V_2)$ can be expanded to

$$E(\bar{x}_m V_2) = u_{\text{chip}} E(V_2) + \mathbf{u_{coeff}} E(\mathbf{z}^T \mathbf{z} V \mathbf{z}^T)$$
$$= u_{\text{chip}} tr(V). \quad (53)$$

Thus, it has been proved that $E(\bar{x}_m)E(v) = E(\bar{x}_m v)$. In other words, $\bar{x}_m$ and v are uncorrelated. ■

References

[1] B. Calhoun, Y. Cao, X. Li, K. Mai, L. Pileggi, R. Rutenbar, and L. Shepard, "Digital circuit design challenges and opportunities in the era of nanoscale CMOS," *Proc. IEEE*, vol. 96, no. 2, pp. 343–365, Feb. 2008.

[2] 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.

[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*, 2006, pp. 18–28.

[4] G. Ribes, J. Mitard, M. Denais, S. Bruyere, F. Monsieur, C. Parthasarathy, E. Vincent, and G. Ghibaudo, "Review on high-k Dielectrics reliability issues," *IEEE Trans. Devices Mater. Reliab.*, vol. 5, no. 1, pp. 5–19, Mar. 2005.

[5] 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.

[6] C. Zhuo, K. Chopra, D. Blaauw, and D. Sylvester, "Process variation and temperature-aware full chip oxide breakdown reliability analysis," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 30, no. 9, pp. 1321–1334, Sep. 2011.

[7] J. Fang and S. Sapatnekar, "Scalable methods for analyzing the circuit failure probability due to gate oxide breakdown," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 11, pp. 1960–1973, Nov. 2012.

[8] International Technology Roadmap for Semiconductors, 2008 Update—Process Integration, Devices, and Structures.

[9] 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.

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

[11] E. Karl, P. Singh, D. Blaauw, and D. Sylvester, "Compact in-situ sensors for monitoring negative-bias-temperature-instability effect and oxide degradation," in *Proc. ISSCC*, 2008, pp. 410–623.
[12] J. Keane, S. Venkatraman, P. Butzen, and C. H. Kim, "An array-based test circuit for fully automated gate dielectric breakdown characterization," in *Proc. CICC*, 2008, pp. 121–124.
[13] X. Li, R. Rutenbar, and R. Blanton, "Virtual probe: A statistically optimal framework for minimum-cost silicon characterization of nanoscale integrated circuits," in *Proc. ICCAD*, 2009, pp. 433–440.
[14] R. Nonnel, S. Thió-Henestrosa, and T. Aluja-Banet, "Some alternatives for conditional principal component analysis," *Appl. Stochastic Models Business Ind.*, vol. 16, no. 2, pp. 147–158, 2000.
[15] C. Zhuo, D. Blaauw, and D. Sylvester, "Post-fabrication measurement-driven oxide breakdown reliability prediction and management," in *Proc. ICCAD*, 2009, pp. 441–448.
[16] J. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Trans. Devices Mater. Reliab.*, vol. 1, no. 1, pp. 43–59, Mar. 2001.
[17] 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.
[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] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Proc. DAC*, 2007, pp. 817–822.
[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] E. Y. Wu, E. Nowak, R.-P. Vollertsen, and L.-K. Han, "Weibull breakdown Characteristics and oxide thickness uniformity," *IEEE Trans. Electron Devices*, vol. 47, no. 12, pp. 2301–2309, Dec. 2000.
[23] R.-P. Vollertsen and W. G. Kleppmann, "Dependence of dielectric time to breakdown distributions on test structure area," in *Proc. ICMTS*, 1991, pp. 75–79.
[24] W. Meeker and L. Escobar, *Statistical Methods for Reliability Data.* New York: Wiley, 1998.
[25] S. Kotz, N. Balakrishnan, and N. Johnson, *Continuous Multivariate Distributions*. New York: Wiley, 2000.
[26] Q. Liu and S. Sapatnekar, "Confidence scalable post-silicon statistical delay prediction under process variation," in *Proc. DAC*, 2007, pp. 496–502.
[27] 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.
[28] 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.
[29] A. Gattiker, M. Bhushan, and M. Ketchen, "Data analysis techniques for CMOS technology characterization and product impact assessment," in *Proc. ITC*, 2006, pp. 1–10.
[30] K.-H. Yuan and P. M. Bentler, "Two simple approximations to the distributions of quadratic forms," *Brit. J. Math. Statist. Psychol.*, vol. 63, no. 2, pp. 273–291, 2010.
[31] 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.
[32] 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 ultrathin gate oxides. *Microelectron. Eng.*, vol. 46, no. 11, pp. 1787–1798, 2002.
[33] R. Degraeve, N. Pangon, B. Kaczer, T. Nigam, G. Groeseneken, and A. Naem, "Temperature acceleration of oxide breakdown and its impact on ultrathin gate oxide reliability," in *Proc. VLSIT*, 1999, pp. 59–60.

Cheng Zhuo (S'06–M'12) 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 University of Michigan, Ann Arbor, in 2010.

He is currently with Intel Corporation, Hillsboro, OR. His current research interests include interconnect and timing analysis, design for reliability, and variability-aware optimization. He has served on technical program committees of ICCAD, ISPD, and ISQED.

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

He is currently a Professor of electrical engineering and computer science with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, and the Director of the Michigan Integrated Circuits Laboratory, a group of ten Faculty Members and over 60 graduate students. He previously held research staff positions with the Advanced Technology Group, Synopsys, Mountain View, CA, Hewlett-Packard Laboratories, Palo Alto, CA, and a visiting professorship in electrical and computer engineering with the National University of Singapore, Singapore. He has published over 300 articles, one book, and several book chapters. He cofounded Ambiq Micro, a fabless semiconductor company developing ultralow power mixed-signal solutions for compact wireless devices. His current research interests include design of millimeter-scale computing systems and energy efficient near-threshold computing for a range of applications. He also serves as a Consultant and Technical Advisory Board Member for electronic design automation and semiconductor firms in these areas. He holds 13 U.S. patents.

Dr. Sylvester was a recipient of 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 was a recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for Distinguished Scholarship. He has served on 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 has served as an Associate Editor for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS.

David Blaauw (M'00–SM'07–F'12) received the B.S. degree in physics and computer science from Duke University, Durham, NC, in 1986, and the Ph.D. degree in computer science from the University of Illinois, Urbana, in 1991.

Until August 2001, he was with Motorola, Inc., Austin, TX, were he was the 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. He has published over 350 papers and holds 40 patents. His work has focused on very large scale integration design with a 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 Electronic and Design. He has also been the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference and a member of the ISSCC Technical Program Committee.