Post-Fabrication Measurement-Driven Oxide Breakdown Reliability Prediction and Management

Cheng Zhuo, David Blaauw, Dennis Sylvester
EECS Department, University of Michigan, Ann Arbor, MI 48109
{czhuo, blaauw, dennis}@eecs.umich.edu

ABSTRACT

Oxide breakdown has become an increasingly pressing reliability issue in modern VLSI design with ultra-thin oxides. The conventional guard-band methodology assumes uniformly thin oxide thickness and results in overly pessimistic reliability estimation that severely degrades the system performance. In this study 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 prediction and quantify the trade-off between reliability margin and system performance. Given the post-fabrication measurements, chip oxide breakdown reliability can be formulated as a conditional distribution that allows us 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 the proposed method can achieve performance improvement of 19% on average and 27% at maximum for a design with up to 50 million devices, using merely 25 measurements per chip, while analysis time is only 0.4 second.

Categories and Subject Descriptors

B.7.2 [**Hardware**]: Integrated Circuits - design aids

General Terms

Performance, Algorithms

Keywords

oxide breakdown, reliability, post-fabrication

1. 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 one 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].

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 strategy is overly pessimistic and enforces an overly low supply voltage for the ensemble of chips, causing significant penalty in performance budget [2, 3]. In practice, no two transistors are exactly the same or have precisely the same characteristics. Instead, they vary significantly from wafer to wafer, reticle to reticle, die to die and across the die. Hence, some dies with thinner than average oxides are much more likely to fail than other dies. To more accurately account for the impact of thickness variation on lifetime prediction, a recent research incorporated both inter- and intra-chip variations into a statistical lifetime analysis [4].

ICCAD'09, November 2–5, 2009, San Jose, California, USA.

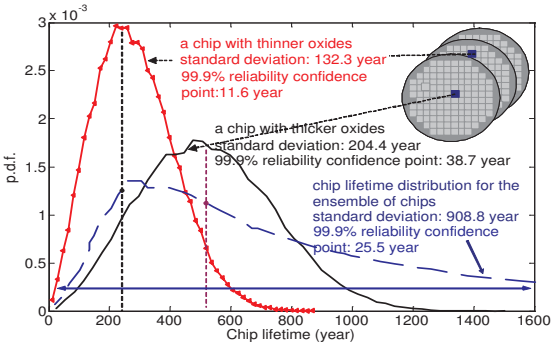

Figure 1: Chip lifetime distribution for the ensemble of chips (blue curve) (nominal thickness is 1.67 nm for 65nm device). Red and black curves represent lifetime distributions for two particular chips with all devices' oxide thicknesses known.

However, without post-fabrication measurement, designers cannot know the oxide thickness of an individual transistor on a particular die and hence cannot determine the specific lifetime expectation from one chip to another. Either the method in [4] or Monte-Carlo simulation only relies on the general variation knowledge of the technology node and results in a more accurate but ultimately still highly-spread lifetime distribution for any chip. This is due in part to the lack of information of the unique condition of a particular chip and unfairly implies a chip that happens to have thicker oxides, bears the same risk to failure as the one with thinner oxides. Figure 1 presents the chip lifetime distribution (blue curve) by simulating the failure time of 50000 chips in a Monte-Carlo fashion. The spread in lifetime results is partly from the innate randomness of the OBD mechanism. The lifetime spread is further increased by thickness variation ($3\sigma/\mu$=4% [5]) which has an exponential effect on the tunnelling current and injected charge and eventually leads to the lognormal shape in Figure 1 with a long tail (908.8 year standard deviation / 99.9% reliability confidence point is 25.5 year) [6]. However, each chip has unique oxide thickness conditions for each transistor and hence some chips are bound to have significant lifetime margin which could be traded off for higher performance by allowing these chips to operate at a higher supply voltage.

Thus, if the oxide thickness of each individual transistor on a fabricated chip could be measured, the lifetime distribution for that chip would be significantly tightened, as shown in Figure 1 for two chips, one with thinner oxides (red curve, 132.3 year standard deviation / 99.9% reliability confidence point is 11.6 year) and one with thicker oxides (black curve, 204.4 year standard deviation / 99.9% reliability confidence point is 38.7 year). Then the chip with thinner oxide thickness (red curve) has a significantly higher risk to fail early and should be operated using a lower maximum supply voltage, thereby improving the overall reliability of the design. Conversely, the chip with black curve whose oxide thickness happens to be thicker is less prone to failure and could be operated at a higher supply voltage limit and therefore obtain a performance gain while still meeting reliability target. Hence, understanding the oxide thickness condition on a die can result in both a performance improvement as well as a higher reliability.

Unfortunately, obtaining oxide thickness condition for all devices on a die is impossible in today's chips with hundreds of millions to billions of transistors. However, recent advances in compact oxide thickness sensors [8, 9] allow tens to hundreds

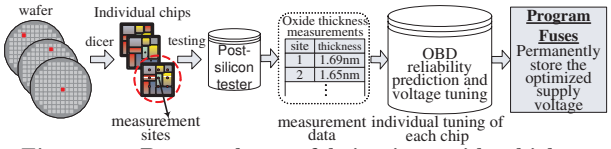

Figure 2: Proposed post-fabrication oxide thickness measurement-driven supply voltage optimization flow.

of sensors to be placed on a chip or even inside cores. Thus, a 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 (<1000) of post-fabrication oxide thickness measurements using on-chip sensors. This problem is non-trivial:

- 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. It is therefore crucial to fully utilize of the measurement information to predict the oxide thickness for all devices as accurately as possible.
- 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 multi-dimensional nested stochastic process. Simple Monte-Carlo simulation must model both the random variation in oxide thickness as well as the innate variation of OBD itself and is therefore extremely expensive in both time and memory.
- Finally, OBD is also a strong function of the chip operating conditions, such as processor state and temperature which vary during the operation of a device. For simplicity, our discussion has not accounted for these factors up to now. However, they have a significant impact on the lifetime of a particular die and we will outline how to incorporate these effects in our proposed analysis.

In this paper, we propose a new *post-fabrication* measurement-driven OBD reliability prediction and management methodology using a limited number of measurement points. The measurements of oxide thicknesses for a subset of devices can be conducted by on-chip sensors [8] or test-structures [9], which can be easily modified to measure the oxide thickness instead of monitoring the degradation process. Figure 2 displays 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 algorithms and, if available, dynamic reliability management algorithms [10] that control the chip operation during runtime.

The OBD reliability prediction and voltage tuning module in this flow consists itself of three phases. The first phase uses limited *post-fabrication* measurements to reduce the uncertainty of the oxide thickness for any unmeasured device. The proposed method accounts for both inter-chip (global), intra-chip (within-chip) spatially correlated and random residual components [7]. We compute the inter-chip component using a maximum-likelihood estimation method and the other two by leveraging the spatial correlation between devices and then constructing a conditional distribution based on the post-fabrication measurements, while still preserving the correlation between devices in a conditional covariance matrix. Based on the conditional distribution, the second phase applies principal component analysis to predict the chip reliability. The 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 certain confidence-level interval, the lower bound of which is conservatively used for lifetime evaluation. Finally, in the third phase, we present an optimization flow for efficient tuning of the chip maximum supply voltage. As a result, with proper reliability management, we can boost chip performance for many chips while maintaining or improving reliability.

2. REVIEW OF OXIDE BREAKDOWN RELIABILITY ANALYSIS

Conventionally, the gate oxide degradation is considered to depend on oxide thickness, transistor area, supply voltage, and temperature. Although many of the physical details are still under debate, most models note the non-deterministic process of defect generation, eventually resulting in a statistically distributed oxide breakdown time and the strong dependence of this random process on oxide thickness [13, 14]. In this section, we gives a brief review of the oxide thickness variation modeling and previous statistical method for OBD reliability analysis.

2.1 Oxide Thickness Variation Modeling

Typically the oxide thickness variation can be classified based on the spatial scale over which it manifests [11, 12]. Given the decomposition of global inter-chip, intra-chip spatially correlated and random variation components, oxide thickness for any device can be modeled as:

$$x = u_0 + z_g + z_{corr} + z_\epsilon \quad (1)$$

where u_0 is the nominal oxide thickness for the technology. z_g denotes the global-scale inter-chip variation component. Clearly, all the devices on the same chip observe the same amount of z_g in oxide thickness, whereas z_g varies for different chips. The fluctuation of z_g among different chips can then be modeled by a Gaussian process $N(0, \sigma_g^2)$ [7]. z_{corr} is the intra-chip spatially correlated component that tends to affect closely-placed devices in a similar manner. A typical modeling of the vector $\mathbf{z_{corr}} = [z_{corr,1}, z_{corr,2}...z_{corr,m}]$ for m devices is a multi-variate Gaussian process, *i.e.*, $\mathbf{z_{corr}} \sim \mathcal{N}_m(0, \Sigma_{corr})$, where the subscript of $\mathcal{N}$ denotes the dimensionality of the random vector, and Σ_{corr} is a $m \times m$ covariance matrix for m devices. A simplified model of spatial correlation can be achieved by partitioning the chip into N grids and assuming perfect spatial correlation within each grid [11, 12]. In other words, the devices within one grid have the correlation coefficient of 1 and hence bear the same spatially-correlated variation component, whereas devices in two different grids, i_{th} and j_{th} for example, have a covariance of $\rho_{i,j}\sigma_{corr}^2$, with a correlation coefficient $\rho_{i,j}$ <1 [11]. Finally, z_ϵ is the random residual variation resulting from certain local device scale effects and is modeled as a Gaussian process $N(0, \sigma_\epsilon^2)$ [7]. In general, σ_g, Σ_{corr} and σ_ϵ denote the uncertainness of the variation components at different spatial-scales, and can be either achieved from prior knowledge or robustly extracted from measurements as in [7, 15].

2.2 Previous Statistical OBD Reliability Analysis

A common failure criterion for OBD is soft breakdown (SBD) characterized by a small increase in gate leakage and eventually followed by un-recoverable hard breakdown (HBD). Due to the stochastic process nature, the oxide breakdown time for SBD is modeled as a random variable following a Weibull probability distribution [13]:

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

where F is the cumulative distribution function (cdf) of time-to-breakdown t, a is the device area normalized with the minimum device area, α and β are the scale and shape parameters of the Weibull model. β can be further expressed as bx for a given temperature and voltage stress, where x is the gate oxide thickness of a device. The reliability function of a device can then be simply written as:

$$R(t) = P(T > t) = 1 - F(t) = e^{-a(\frac{t}{\alpha})^{bx}} \quad (3)$$

Due to the non-deterministic characteristic of oxide thickness at design time, the device reliability function can be in-

terpreted as the conditional reliability function given its oxide thickness and written as $R(t|x_i)$. The overall chip-level reliability function is then given by:

$$R_c(t) = \int_0^\infty \dots \int_0^\infty \prod_{i=1}^{m} R_i(t|x_i) f(x_1 \dots x_m) dx_1 \dots dx_m \quad (4)$$

where $f(x_1, \dots, x_m)$ is the joint probability density function (pdf) of the gate oxide thicknesses for m devices. To handle the tremendous dimensionality of (4), [4] proposed to project the parametric space to two distinct random variables, sample mean (u) and variance (v) of the chip oxide thickness distribution. Based on this, the original product $\prod_{i=1}^{m} R_i(t|x_i)$ was simplified to a conditional probability $R_c(t|u,v)$ [4]. The integral of (4) is then compactly expressed as:

$$R_c(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_c(t|u,v) f_{\mathbf{uv}}(u,v) du\ dv \quad (5)$$

where

$$R_c(t|u,v) = \exp[-Ae^{\ln(\frac{t}{\alpha})bu + (\ln(\frac{t}{\alpha}))^2 b^2 v/2}] \quad (6)$$

and $f_{\mathbf{uv}}(u,v)$ is the joint pdf of a Gaussian random variable u and a chi-square random variable v.

However, neither the method in [4] nor the guard-band method in [3] allows for incorporation of oxide thickness measurements and is unable to distinguish the unique condition of a particular chip. These methods therefore result in one global lifetime estimation for the entire ensemble of chips, and unnecessarily degrade the performance for most of them.

3. POST-FABRICATION MEASUREMENT-DRIVEN OXIDE THICKNESS ESTIMATION

We will show that even with a relatively small number of oxide thickness measurements, it is possible to reduce the uncertainty of oxide thicknesses for a particular chip, and hence provide significantly more accurate lifetime estimation. However, due to the tremendous number of unmeasured devices and the constrained stochastic process nature of chip reliability, such estimation of oxide thicknesses for unmeasured devices is a difficult problem that has not been addressed to date. This section presents a statistical method to address this problem.

3.1 Problem Formulation

Give a *particular* chip, the inter-chip and intra-chip variation components (spatially correlated and random) play very different roles in the final transistor oxide thickness. The inter-chip component induces the same increment or decrement to the oxide thicknesses for all the devices within the chip and is a constant in (1). On the other hand, the intra-chip spatially correlated and random components are different from device to device. In reality we cannot distinguish the sources of the variation when the number of measurements is limited. Thus, we combine the two intra-chip variation components together in analysis and comprehensively evaluate their impact.

Given a chip dissected to N grids as in [11] with m devices in total, the vector of oxide thicknesses for all the devices can be written as:

$$\mathbf{x} = u_0 + z_g + \mathbf{z_{corr}} + \mathbf{z}_\epsilon = u_{chip} + \mathbf{z_{intra}} \quad (7)$$

where $\mathbf{x}$=$[x_1, x_2, \dots, x_m]$ 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; $\mathbf{z}_\epsilon$ is the vector containing the random variation component of each device; $\mathbf{z_{intra}}$=$\mathbf{z_{corr}}$+$\mathbf{z}_\epsilon$ is hence the combined intra-chip variation component that preserves the spatial correlation between devices. Since $\mathbf{z}_\epsilon$ can be interpreted as a multi-variate 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 then the sum of two multi-variate Gaussians and remains a multi-variate Gaussian process $\mathcal{N}_m(0, \Sigma_{intra})$, where $\Sigma_{intra} = \Sigma_{corr} + \sigma_\epsilon^2 I_m$.

The post-fabrication measurement-driven oxide thickness estimation problem is then formulated as:

Formulation: *Given the thickness variation model in* (7) *and the oxide thickness measurements of n_0 devices across a particular chip, estimate the oxide thickness of any unmeasured device, including the components of u_{chip} and $\mathbf{z_{intra}}$ as well as the corresponding variance.*

In the following, we present the techniques to solve the above formulation.

3.2 Model Simplification

The grid-based spatial correlation model in [11] indicates that devices within one grid bear approximately the same inter-chip and intra-chip spatially correlated variation components. This is reasonable when we have relatively finer grids across the chip. The difference in 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 predicted. Thus, instead of performing device-level estimation and predicting device by device within one grid, we employ a grid-based prediction scheme by associating every grid with one random variable and hence achieve one estimation for each grid, including a random variation component and correlation to other grids. Clearly, such modeling greatly simplifies the complexity from the dimensionality of millions (number of devices) to $N + n_0$, where n_0 is the number of measurement sites and N denotes the number of unmeasured sites with each representing one grid.

We then re-formulate the model in (7) to the granularity of a grid. Both $\mathbf{x}$ and $\mathbf{z_{intra}}$ are now $(N+n_0)\times 1$ vectors. $\mathbf{z_{intra}}$ follows $\mathcal{N}_{N+n_0}(0, \Sigma_{intra,grid})$, where $\Sigma_{intra,grid}$ is an $(N+n_0)\times(N+n_0)$ covariance matrix for N unmeasured sites and n_0 measured sites.

3.3 Estimation of the Chip-Level Oxide Thickness Mean u_{chip}

As discussed, we need to treat the deterministic component and random component in (7) separately. Removal of the mean from the random data is an integral and essential step to minimize the mean square error of the estimation [16]. In this subsection, we detail the estimation of the chip-level oxide thickness mean u_{chip}.

Before measurement, the oxide thickness for the sites to be measured remain unknown and hence can be characterized by a multi-variate Gaussian model, $\mathcal{N}_{n_0}(u_{chip}, \Sigma_{mm})$. The measured thicknesses $\mathbf{s}$=$[s_1, s_2 ... s_{n_0}]$ are therefore a *sample vector* drawn from this stochastic model, with measurements acting as n_0 observations. Thus, by using the maximum likelihood estimation (MLE), the log-likelihood function is [16]:

$$\begin{aligned} \ell(\mathbf{s}|u_{chip}) &= -\ln((2\pi)^{n_0/2}|\Sigma_{mm}|^{1/2}) \\ &- \frac{1}{2}(\mathbf{s} - u_{chip} \times [\mathbf{1}]_{1\times n_0})\Sigma_{mm}^{-1}(\mathbf{s} - u_{chip} \times [\mathbf{1}]_{1\times n_0})^T \end{aligned} \quad (8)$$

where $[\mathbf{1}]_{1\times n_0}$ denotes a $1\times n_0$ all-one vector. The maximum in (8) is achieved when:

$$u_{chip} \approx \frac{[\mathbf{1}]_{1\times n_0}\Sigma_{mm}^{-1}}{[\mathbf{1}]_{1\times n_0}\Sigma_{mm}^{-1}[\mathbf{1}]_{1\times n_0}^T}\mathbf{s}^T \quad (9)$$

The corresponding MLE estimation variance can be approximately bounded by the Cramer-Rao bound [16]:

$$\text{var}(u_{chip}) \approx [\mathbf{1}]_{1\times n_0}\Sigma_{mm}^{-1}[\mathbf{1}]_{1\times n_0}^T \quad (10)$$

Since the number of measurements n_0 is limited to fewer than hundreds, the matrix inverse in (9) can be efficiently computed within seconds.

3.4 Estimation of the Intra-Chip Variation Component $\mathbf{z_{intra}}$

If every site of a chip could be measured, the variance for the random vector $\mathbf{x}$ would be reduced to 0. Since the number of measurements is limited, measured oxide thicknesses can only reduce the variance of unmeasured sites, which can still provide designers with significantly more accurate information of chip oxides condition.

In order to assess the impact of measurements, we separate the oxide thickness vector $\mathbf{x}$ into two sub-vectors 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. $\Sigma_{intra,grid}$ then can be expressed as:

$$\Sigma_{intra,grid} = \begin{bmatrix} \Sigma_{mm} & \Sigma_{mu} \\ \Sigma_{um} & \Sigma_{uu} \end{bmatrix} \quad (11)$$

where in each sub-matrix, "m" is for the sites to be measured (vector "$\mathbf{s}$") and "u" is for the unmeasured sites (vector "$\mathbf{x_u}$"). The entries in any sub-matrix can be simply obtained from

the covariance matrix Σ_{intra} in (7) by identifying the grids the sites belong to. Note that both $\mathbf{s}$ and $\mathbf{x_u}$ are multi-variate Gaussian processes with a mean of u_{chip} and a covariance matrices of Σ_{mm} and Σ_{uu}, respectively.

Given the measurement values $\mathbf{s} = \mathbf{s_0}$ at n_0 sites, the sub-vector $\mathbf{x_u}$ 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 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:

$$f_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}(\mathbf{x_u}) = \frac{f_{\mathbf{x}}(\mathbf{x_u}, \mathbf{s} = \mathbf{s_0})}{f_{\mathbf{s}}(\mathbf{s} = \mathbf{s_0})} \tag{12}$$

where $f_{\mathbf{x}}(\mathbf{x})$ and $f_{\mathbf{s}}(\mathbf{s})$ are pdf's for the multi-variate 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 of the deduction.

Based on the decomposition of the covariance matrix in (11), we define:

$$\mathbf{u}_{\mathbf{x_u}|\mathbf{s}} = u_{chip} + (\mathbf{s} - u_{chip})\Sigma_{mm}^{-1}\Sigma_{mu} \tag{13}$$

$$\Sigma_{\mathbf{x_u}|\mathbf{s}} = \Sigma_{uu} - \Sigma_{um}\Sigma_{mm}^{-1}\Sigma_{mu} \tag{14}$$

and then obtain $|\Sigma_{intra,grid}| = |\Sigma_{mm}||\Sigma_{\mathbf{x_u}|\mathbf{s}}|$. Thus, when $\mathbf{s} = \mathbf{s_0}$, the conditional pdf in (12) can be expressed as:

$$f_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}(\mathbf{x_u}) = \frac{1}{(2\pi)^{N/2}|\Sigma_{\mathbf{x_u}|\mathbf{s}}|^{1/2}} \times \exp[\frac{-(\mathbf{x_u} - \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}})\Sigma_{\mathbf{x_u}|\mathbf{s}}^{-1}(\mathbf{x_u} - \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}})^T}{2}] \tag{15}$$

where $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ and $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ defined in (13) and (14) are conditional mean and covariance matrix for the conditioned random vector $\mathbf{x_u}|_{\mathbf{s}=\mathbf{s_0}}$. The details of the conditional distribution can be derived from general principles in [17], which are widely employed in various works [18, 19].

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. Note that every entry in the covariance matrix is positive and that both Σ_{uu} and Σ_{mm} are positive definite [7]. The conditional variance in $\Sigma_{\mathbf{x_u}|\mathbf{s}}$ is therefore reduced compared with the unconditional variance in (11).

Although the oxide thicknesses for closely-placed devices are non-continuous due to random variation, the spatial correlation still allows us to explore the relationship among devices and achieves improved prediction as the number of measurements increases. Figure 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 with respect to the growing number of measurements. It is noted that with only 9 measurements, the variance of $u_{x_u,i}|_{\mathbf{s}=\mathbf{s_0}}$, as computed in (14), is reduced by 63% compared with the initial variance when no measurement is conducted.

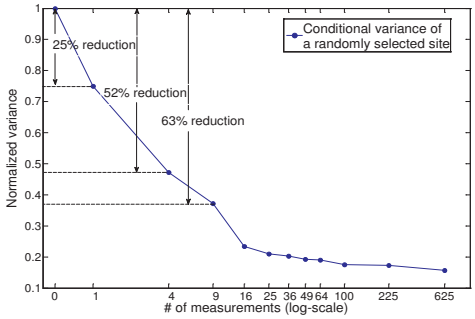

Figure 3: Reduction in variance of the conditional estimator $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ for a randomly selected unmeasured site with increased measurements.

3.5 Chip-Level Oxide Thickness Mean Refinement

Unlike the unconditional random vector in (7) where all the variables share the same mean u_{chip}, the conditioned random vector in (13) may bear different mean values. This is closer to the realistic condition where the oxide thickness shows variation across the die, and hence provides the chance to refine the chip-level oxide thickness mean, using both the measured and unmeasured sites.

In theory, the chip-level oxide thickness mean is equal to the sample mean of all the sites, denoted as $\overline{x}_{N+n_0}$:

$$\overline{x}_{N+n_0} = \frac{1}{N+n_0}(\mathbf{s_0} \times [\mathbf{1}]_{n_0 \times 1} + \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} \times [\mathbf{1}]_{N \times 1}) \tag{16}$$

The deviation between $\overline{x}_{N+n_0}$ and u_{chip} is primarily due to the estimation error and may degrade the analysis effectiveness. Thus, we can perform a refinement step iteratively to reduce the deviation to a negligible level, *i.e.*, to make $u_{chip} \approx \overline{x}_{N+n_0}$ by repeatedly replacing u_{chip} in (13) with $\overline{x}_{N+n_0}$ and then computing $\overline{x}_{N+n_0}$ with (16). In general the refinement is completed within tens of iterations to reach certain tolerance, *e.g.*, 10^{-5}. Moreover, it is worthwhile to note that either the estimation variance in (10) or the conditional covariance matrix in (14) does not rely on u_{chip} and remains unchanged for the updated chip-level oxide thickness mean.

Procedure: *Post-Fabrication Measurement-Driven Estimation*	
Input: measurements $\mathbf{s_0}$, process variation model in (7) **Output:** Oxide thickness estimation for each device and the corresponding estimation variance	
1:	Simplify the model as in Subsection 3.2;
2:	Compute the chip-level oxide thickness mean and corresponding variance using (9) and (10);
3:	Estimate the intra-chip variation component $\mathbf{z_{intra}}$ using (13)-(15);
4:	Perform chip-level oxide thickness mean refinement;
5:	Map the estimation and corresponding variance at the granularity of grid level to the devices in the same grid;

Figure 4: Post-fabrication measurement-driven oxide thickness estimation.

3.6 Summary of Post-Fabrication Measurement-Driven Estimation

We summarize the procedure for post-fabrication measurement-driven estimation in Figure 4. Note that the procedure produces a single random value per grid which is the representative for all the devices in the grid. The estimation for this site is then eventually projected to all the other devices within the same grid to compute the reliability of the chip.

We apply the proposed procedure to 10000 chips in 65nm technology. Each chip has 0.5 million devices and is imposed a 50×50 (=2500) grids with 100 uniformly-distributed measurement sites. The estimated chip-level oxide thickness mean u_{chip} is compared with the actual mean of the oxide thicknesses for all the devices in Figure 5. From either the histogram or the scatter plot, it can be seen that the estimation achieved by maximum likelihood estimation (MLE) in Subsection 3.3 is very accurate with a maximum relative error of 0.77% while the mean refinement algorithm (in Subsection 3.5) can further reduce the relative error to 0.33%. We then examine the estimation accuracy at the device level (achieved by step 5 in Figure 4) for a randomly selected chip. Figure 6 demonstrates the contour of the difference between the actual oxide thickness and the estimated thickness mapped from $\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}}$ for all the devices on a chip. One can see that 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 $\Sigma_{\mathbf{x_u}|\mathbf{s}}$.

4. MEASUREMENT-DRIVEN OBD RELIABILITY PREDICTION AND MANAGEMENT

Using the proposed oxide thickness estimation with corresponding variance, we now perform a statistical reliability analysis to tighten the lifetime distribution. Here we focus on chip-level reliability analysis and consider the worst-case operating temperature to ensure a correct operation throughout

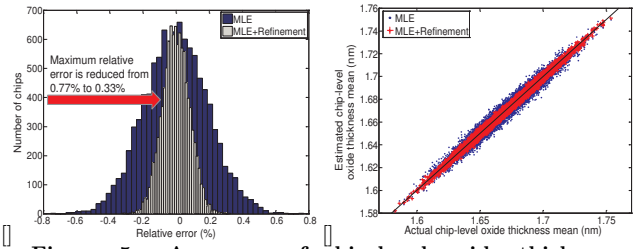

Figure 5: Accuracy of chip-level oxide thickness mean estimation: (a) Histograms of relative errors for maximum likelihood estimation (MLE) and maximum likelihood estimation with refinement (MLE+Refinement) (b) Scatter plots for MLE and MLE+Refinement.

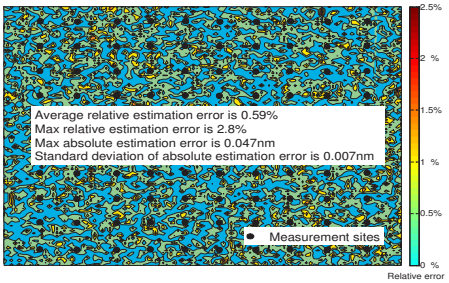

Figure 6: Contour of the device-level oxide thickness estimation error for a chip with 0.5M devices and 100 measurements [1].

the entire lifetime. The temperature and voltage drop variations can be easily incorporated in our flow by performing analysis at the granularity of functional blocks or sub-blocks, where devices within a block can be assumed to have the same temperature and supply voltage drop.

Given a chip with m devices and N grid cells for spatial correlation modeling, we define the following notations for the remainder of the paper as in Table 1.

4.1 Post-Fabrication Measurement-Driven Reliability Prediction

The challenge to the chip-level statistical OBD reliability analysis is the huge dimensionality of the integral in (4). Reference [4] 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 [4]. The conditional covariance matrix (or spatial correlation) also shows completely different features from the unconditional case. We therefore present a measurement-driven OBD reliability analysis in this subsection.

4.1.1 *Conditional Spatial Correlation Characterization Using Principal Component Analysis*

Since $\mathbf{x_u}|\mathbf{s_0}$ is still a multi-variate Gaussian random vector, its correlation structure in (14) can be simplified by principal component analysis (PCA) to map the correlated variation components to another set of mutually independent random variables with zero mean and unit variance [4, 11]. For a device in the i_{th} grid, its conditional oxide thickness $x_{u,i}|\mathbf{s_0}$ can be canonically expressed as a linear combination of the principal components:

$$x_{u,i}|\mathbf{s_0} = u_{x_{u,i}|\mathbf{s}=\mathbf{s_0}} + \sum_{j=1}^{N} \lambda_{i,j} z_j \tag{17}$$

where N is the number of principal components; z_j's represent the N independent random variables used to characterize

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

[2] 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.

Table 1: Notations used in OBD reliability analysis

Notation	Definition
$\mathrm{x}=[\mathrm{x}_1,\ldots,x_m]$	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 $\mathbf{s_0}$
$\overline{x}_m=\frac{\sum_{i=1}^{m} x_i}{m}$	the sample mean for m devices of a chip
$v=\frac{\sum_{i=1}^{m}(x_i-\overline{x}_m)^2}{m-1}$	the sample variance for m device 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)[2], 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

the conditional spatially correlated variation; and the coefficients $\lambda_{i,j}$'s represent the sensitivity of thickness variation with respect to the j_{th} principal component for the random variable in the i_{th} grid. Thus, the conditional random vector of N unmeasured sites can be written compactly with principal components:

$$\mathbf{x_u}|\mathbf{s_0} = \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} + \mathbf{z} \times P_\lambda \tag{18}$$

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; $\mathbf{z} = [z_1, z_2, \ldots z_N]$ is a vector of principal components.

We can then estimate the *conditional* sample mean and sample variance for devices across the chip in terms of principal components. As defined earlier, $\overline{x}_m$ and v are:

$$\overline{x}_m = [(\mathbf{x_u}|\mathbf{s_0})D_0^T + \mathbf{s_0} \times [\mathbf{1}]_{n_0 \times 1}]/m \tag{19}$$

$$v = \frac{(\mathbf{x_u}|\mathbf{s_0} - \overline{x}_m)D(\mathbf{x_u}|\mathbf{s_0} - \overline{x}_m)^T + (\mathbf{s_0} - \overline{x}_m)(\mathbf{s_0} - \overline{x}_m)^T}{m-1} \tag{20}$$

Those two variables $\overline{x}_m$ and v illustrate the underlying characteristics of the conditional chip oxide thickness distribution given measurements $\mathbf{s_0}$ at n_0 sites.

By noting the equality of u_{chip} in (16), (19) is simplified to:

$$\overline{x}_m = u_{chip} + \mathbf{u_{coeff}}\mathbf{z^T} \tag{21}$$

where $\mathbf{u_{coeff}} = \frac{1}{m} D_0 P_\lambda^T$. Clearly $\overline{x}_m$ remains a Gaussian with mean of u_{chip} and variance as the following:

$$\mathrm{var}(\overline{x}_m) = \mathrm{var}(u_{chip}) + \mathbf{u_{coeff}}\mathbf{u_{coeff}}^T \tag{22}$$

After expanding the numerator of (21), we can re-write v as the sum of two random variables V_1 and V_2:

$$v = (V_0 + 2V_1 + V_2)/(m-1) \tag{23}$$

where $V_0 = (\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} - u_{chip})D(\mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} - u_{chip})^T + (\mathbf{s_0} - u_{chip})(\mathbf{s_0} - u_{chip})^T$ (24)

$$V_1 = \mathbf{v_{coeff}}\mathbf{z}^T \quad \text{and} \quad V_2 = \mathbf{z}V\mathbf{z}^T$$

with $\mathbf{v_{coeff}} = \mathbf{u}_{\mathbf{x_u}|\mathbf{s}=\mathbf{s_0}} D P_\lambda^T - (m \times u_{chip} - \mathbf{s_0}[\mathbf{1}]_{n_0 \times 1})\mathbf{u_{coeff}}$ and $V = (P_\lambda^T + [\mathbf{1}]_{N\times 1}\mathbf{u_{coeff}})^T D (P_\lambda^T - [\mathbf{1}]_{N\times 1}\mathbf{u_{coeff}})$.

Note that V_0 is a constant and V_1 is a normal random variable. Since the matrix V is positive and symmetric, V_2 has the form of quadratic normal product and can be approximated by a chi-square distribution [21], $V_2 \sim \hat{a}\chi^2_{\hat{b}}$, with $\hat{a} = \frac{tr(V^2)}{tr(V)}$ and $\hat{b} = \frac{[tr(V)]^2}{tr(V^2)}$, where $tr[.]$ denotes the trace operation to compute the sum of diagonal entries. Since $E(V_1)E(V_2)$=$E(V_1V_2)$, V_1 and V_2 turn out to be uncorrelated. Moreover, by noting the degree of freedom for the chi-square distribution $\hat{b} = \frac{[tr(V)]^2}{tr(V^2)}$ is close to N, the chi-square distribution with a large degree of freedom can be well approximated by a Gaussian distribution [16], which is validated by the histogram of V_2 in Figure 7. Thus, the un-correlation between two Gaussian random variables V_1 and V_2 implies their independence. In other words, v is a Gaussian random variable, the mean and variance of which can be computed from (23) and (24):

$$E(v) = [V_0 + tr(V)]/(m-1) \tag{25}$$

$$\mathrm{var}(v) = \frac{2tr(V^2)}{(m-1)^2} + \frac{4}{(m-1)^2}\mathbf{v_{coeff}}\mathbf{v_{coeff}}^T \tag{26}$$

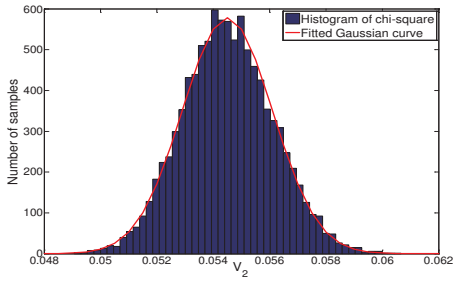

Figure 7: Comparison of the histogram of chi-square random variable V_2 in (24) with degree of freedom of 2209 (N=2500) and the fitted Gaussian curve. The fitting goodness is 0.98 (R-square).

4.1.2 Post-Fabrication Measurement-Driven Lifetime Prediction

Once the underlying distribution 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 (QTTF) [3] for certain reliability target R_t by using (6):

$$R(T_q|\overline{x}_m, v) = \exp[-Ae^{\ln(\frac{T_q}{\alpha})b\overline{x}_m + (\ln(\frac{T_q}{\alpha})b)^2 v/2}] = R_t \quad (27)$$

where A is the chip area. This equality illustrates the actual quantile-based time-to-failure when chip meets ceratin reliability target. The quantile-based time-to-failure 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:

$$v/2 \times \gamma^2 + \overline{x}_m \times \gamma - \ln(-\ln(R_t)/A) = 0 \quad (28)$$

This quadratic equation can be easily solved:

$$\gamma = \gamma(\overline{x}_m, v) = \frac{-\overline{x}_m + \sqrt{\overline{x}_m^2 + 2\ln(-\ln(R_t)/A) \times v}}{v} \quad (29)$$

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

By noting that both $\overline{x}_m$ and v have limited variance, we can further simplify (29) with first-order Taylor expansion:

$$\gamma \approx \gamma(E(\overline{x}_m), E(v)) + [\frac{\partial\gamma(\overline{x}_m, v)}{\partial\overline{x}_m}, \frac{\partial\gamma(\overline{x}_m, v)}{\partial v}]|_{E(\overline{x}_m), E(v)} \times [\overline{x}_m - E(\overline{x}_m), v - E(v)]^T \quad (30)$$

Since both $\overline{x}_m$ and v are Gaussians and almost uncorrelated, we can reasonably justify that γ follows a Gaussian process with mean and variance:

$$E(\gamma) = \frac{-E(\overline{x}_m) + \sqrt{E(\overline{x}_m)^2 + 2\ln(-\ln(R_t)/A) \times E(v)}}{E(v)} \quad (31)$$

$$\text{var}(\gamma) = [(\frac{\partial\gamma}{\partial\overline{x}_m})^2, (\frac{\partial\gamma}{\partial v})^2]|_{E(\overline{x}_m), E(v)} \times [\text{var}(\overline{x}_m), \text{var}(v)]^T \quad (32)$$

Quantile-based time-to-failure T_q can then be characterized as a lognormal distribution, as $T_q = \alpha\exp[\gamma/b]$.

4.2 OBD Reliability Management and Performance Optimization

The technique in Subsection 4.1 can characterize the distribution of quantile-based time-to-failure and achieve a well-tightened lifetime distribution. 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 failure may not exceed $1-R_t$ within T_{target} years lifetime. However, due to the process variation, some chips will have thinner oxides and are quicker to fail. The tightened distribution of T_q (QTTF) enables us to quantitatively evaluate whether the chip will meet the design lifetime target or not. Those chips that are prone to failure can be tuned to a lower supply voltage limit to improve the reliability yield. On the other hand, chips with thicker oxides can operate at a higher voltage for better performance. 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.

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

Procedure: *Post-Fabrication Measurement-Driven OBD Reliability Prediction and Management*	
Input: measurements $\mathbf{s_0}$, process variation model in (7), reliability target and design lifetime	
Output: optimized supply voltage	
1:	Given $\mathbf{s_0}$, estimate the conditional oxide thickness and covariance matrix with the flow in Figure 4;
2:	Apply PCA to the conditional covariance matrix to obtain the distributions of $\overline{x}_m$ and v using (21)-(26);
3:	Estimate tightened chip lifetime distribution using (31) and (32);
4:	Solve the optimization problem in (34)-(36) to achieve the optimized supply voltage;

Figure 8: Post-fabrication measurement-driven OBD reliability prediction and management.

Since QTTF itself is a distribution due to the remaining uncertainty of the oxide thicknesses, we use the lower bound of the distribution with a certain confidence to ensure a robust design. Conservatively, with a 99.9% confidence level, we can derive the following one-sided confidence interval:

$$T_q \in [\alpha\exp\left[\frac{E(\gamma) - 3\sqrt{\text{var}(\gamma)}}{b}\right], \infty] \quad (33)$$

where the moments of γ can be computed from (31) and (32). The lower bound of (33) 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 (33) depend on supply voltage, we formulate the following to maximize the supply voltage while T_{lb} meets the design lifetime target:

$$\textbf{Maximize} \quad v_{chip} \quad (34)$$

Subject to:

$$\ln(T_{lb}) = \ln(\alpha(v_{chip})) + \frac{E(\gamma) - 3\sqrt{\text{var}(\gamma)}}{b(v_{chip})} \geq \ln(T_{target}) \quad (35)$$

$$v_{min} \leq v_{chip} \leq v_{max} \quad (36)$$

where v_{chip} denotes the supply voltage; the first constraint in (35) implies that the 99.9% confidence lower bound of QTTF is larger than the design lifetime target; and the second constraint in (36) denotes the possible voltage tuning range. We find that this optimization problem is equivalent to finding the feasible domain of the inequality in (35), where the parameters of the device reliability function, $\alpha(v_{chip})$ and $b(v_{chip})$, indicate the underlying dependence on supply voltage. Since we only have one variable, even with a complicated physics-based model for $\alpha(v_{chip})$ and $b(v_{chip})$, we can still efficiently solve this problem in a numerical way. In our implementation we adopt the linear models in [13, 14], *i.e.*, $\ln(\alpha(v_{chip}))$=$a_1 \times v_{chip}$+a_2 and $b(v_{chip}) = b_1 \times v_{chip} + b_2$, and hence have a quadratic inequality in (35), which can be analytically computed. As a result, the optimization flow above eventually reduces the failure rate to improve reliability yield, while the overall performance is also enhanced by reducing lifetime safety margins.

4.3 Summary of OBD Reliability Prediction and Management

The procedure for post-fabrication measurement-driven reliability prediction and management is summarized in Figure 8. Given n_0 measurements for a particular chip, we first estimate the oxide thicknesses and corresponding variance using a conditional multi-variate Gaussian model. The conditional spatial correlation is then explored by 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 trade-offs between reliability and supply voltage/performance.

5. EXPERIMENTAL RESULTS

The proposed reliability prediction and management methodology was implemented and tested on several designs using

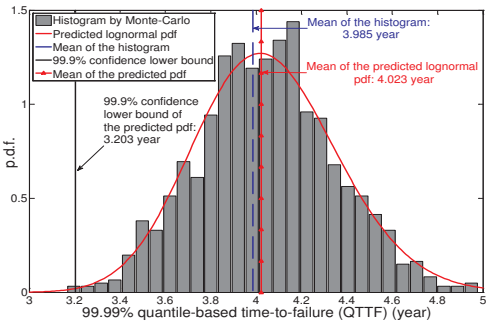

Figure 9: Accuracy comparison of the quantile-based time-to-failure (QTTF) histogram generated by Monte-Carlo simulation and the predicted QTTF pdf using the proposed method. The reliability target R_t is set to 99.99% (100 failures per million).

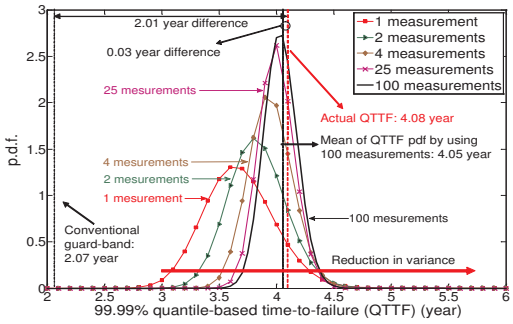

Figure 10: Reduction in the variance of 99.99% quantile-based time-to-failure (QTTF) distribution for a particular chip with increased measurements (1, 2, 4, 25 and 100 measurements).

65nm LP devices (nominal oxide thickness is 1.67nm). The defect generation relationships for the technology node and the technology dependent parameters of the oxide reliability function model are obtained from [13, 14]. In practice, this can be obtained by a one time per technology characterization using test devices [9]. For each design, we used 10000 chips that follow the thickness variation model in Section II. The overall $3\sigma/\mu$ of oxide-thickness variation was set to 4% of the nominal value as in [5] and then split into three variation components.

5.1 Efficacy of the Proposed OBD Reliability Prediction

Given the post-layout design implementation, a process variation model and limited measurements on device oxide thickness, the proposed method can estimate the quantile-based time-to-failure (QTTF) distribution for a certain reliability target, with which we can examine whether this chip may meet the design lifetime or not. 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. In other words, the simulation only accepted the sample vectors with similar entries on the measurement sites, the tolerance of which was set to 0.01nm in our implementation. The results are shown in Figure 9 for a chip with 0.5 million devices and 25 measurements. It is clear that the histogram of 1000 sample vectors matches well with the predicted lognormal pdf using the techniques in Section III and IV. 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 year demonstrating the tightness of the QTTF distribution.

We also explored how the predicted QTTF distribution changes when we increase the number of measurements. Figure 10 clearly shows the reduction in variance as the number of measurements grows. It is interesting to note that even one or two measurements provide sufficient information to tighten the distribution whereas 100 measurements help reduce the standard deviation of the distribution to only 0.16 year. The difference between the actual QTTF and the mean of the predicted QTTF distribution (using 100 measurements) is only 0.03 year (0.8% estimation error), while the conventional guard-band is 2.07 year with almost 50% estimation error.

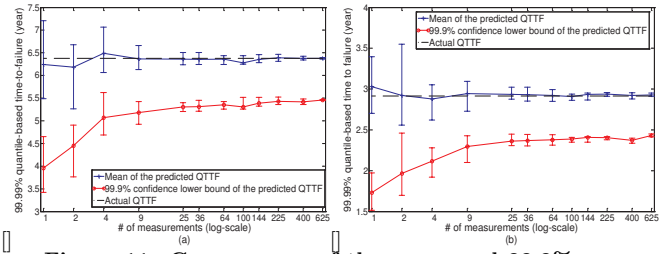

Figure 11: Convergence of the mean and 99.9% confidence lower bound of the predicted quantile-based time-to-failure (QTTF) distribution with increased measurements: (a) a chip with thicker oxides (b) a chip with thinner oxides.

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 the error-bar plots of Figure 11 for two chips, one with thicker oxides and another with thinner oxides. For each particular measurement number, we picked up 10 different configurations (placement) of measurement sites and then computed the 10 set mean/ 99.9% confidence lower bound of QTTF distribution to achieve the error bar. With an increasing number of measurements, both the estimated values and their variance converge quickly.

5.2 Reliability Management and Performance Optimization

Finally, we applied the proposed post-fabrication measurement-driven methodology to tune the supply voltage of 10000 chips of a 0.5M-device design to ensure reliability while maximizing performance. The lifetime target was set to 4 years and the supply voltage tuning range is 0.8V-1.3V.

Figure 12 displays the tuning results using a conventional guard-band, the statistical analysis in [4] (denoted as 0 measurement in the figure) and the proposed methodology using different number of measurements. The guard-band that assumes minimum oxide thickness across the chip, achieved a single supply voltage for all the chips (0.858V) 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 [4] 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 0.875V. 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 measurements, the proposed methodology can obtain a well-tightened QTTF distributions and a more precisely optimize voltage for each chip, achieving 15% performance improvement on average and 26% improvement at maximum. Moreover, although [4] predicts chip lifetime with 99.9% confidence lower bound, still 12 out of 10000 chips fail to meet the lifetime target after tuning, which is beyond the confidence interval. Meanwhile, since the proposed methodology provides more accurate prediction it quickly reduces the number of failures to 0 out of 10000 with increased measurements.

Figure 13 presents the distributions of optimized supply voltage and the resulting performance improvement using different numbers of measurements in tuning. Both the plots show a shift to the right with increased measurements, indicating the capability of the proposed method to choose a more reasonable supply voltage when the number of measurements is increased.

The scalability of the proposed methodology is examined in both its dependence on design complexity/size and run time. 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 cor-

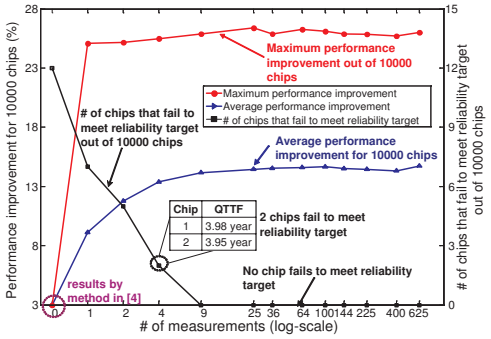

Figure 12: Reliability management results with increased measurements for 10000 chips of a 0.5M-device design: performance improvement and number of tuned chips that fail to meet target after optimization.

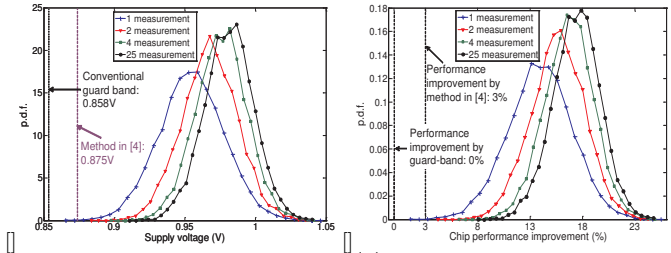

Figure 13: Distributions of (a) optimized supply voltages and (b) performance improvement with different numbers of measurements.

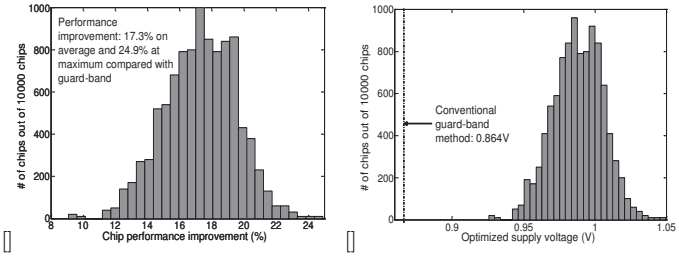

Figure 14: (a) Performance improvement histogram and (b) optimized supply voltage histogram of 10000 chips for an alpha-processor-like design with 0.84M devices and 25 measurements.

relation model have non-uniform densities, i, e, each grid has different number of devices. We measured 25 devices per chip and tuned 10000 chips resulting in a performance improvement of 24.9% at maximum and 17.3% on average compared with conventional guard-band, as shown in Figure 14. We then applied the proposed method to tune 10000 chips of seven different designs (varying in size from 80K to 50M devices) with 25 measurements for each and recorded performance improvement and average run time per chip. Figure 15 shows a flat curve of runtime of around 0.38 second, and a slightly growing trend of average performance improvement from 15% to 19% and maximum improvement from 22% to 27%. As stated earlier, both PCA and matrix inverse are performed once for one design with fixed measurement sites, whereas the analysis and optimization are mostly analytically achievable. Thus, the methodology runtime only relies 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.

6. CONCLUSIONS

This paper presents a post-fabrication measurement-driven OBD reliability prediction and management methodology. The methodology uses limited measurements to estimate the oxides condition of a chip. The estimation is then incorporated into a statistical model to predict a more accurate chip lifetime

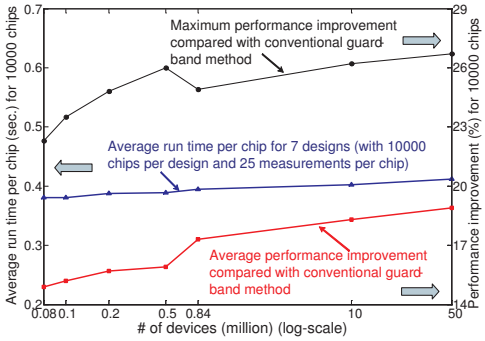

Figure 15: Average run time per chip and average performance improvement for seven different-sized designs (10000 chips for each design and 25 measurements per chip).

distribution, which is fed to an optimization flow to trade off reliability margin and system performance. Experimental results show 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 average run time is only 0.4 second.

7. ACKNOWLEDGEMENT

The authors gratefully acknowledge the National Science Foundation (NSF) for supporting this work.

8. REFERENCES

[1] B. Calhoun, *et.al.* Digital circuit design challenges and opportunities in the era of nanoscale CMOS. *Proceedings of the IEEE (PTI)*, vol. 96, no. 2:343–365, 2008.

[2] S. Borkar. Designing reliable systems from unreliable components: the challenges of transistor variability and degradation. *IEEE Micro.*, vol. 25, issue 6:10–16, 2005.

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

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

[5] International technology roadmap for semiconductors, 2008 update -process integration, devices, and structures.

[6] E. Karl, *et.al.* Analysis of system-level reliability factors and implications on real-time monitoring methods for oxide breakdown device failures. In *Proc. ISQED*, pages 391–395, 2008.

[7] J. Xiong, *et.al.* Robust extraction of spatial correlation. In *Proc. ISPD*, pages 2–9, 2006.

[8] E. Karl, *et.al.* Compact in-situ sensors for monitoring negative-bias-temperature-instability effect and oxide degradation. In *Proc. ISSCC*, pages 410–623, 2008.

[9] J. Keane, *et.al.* An array-based test circuit for fully automated gate dielectric breakdown characterization. In *Proc. CICC*, pages 121–124, 2008.

[10] E. Karl, *et.al.* Reliability modeling and management in dynamic microprocessor-based systems. In *Proc. DAC*, pages 1057–1060, 2006.

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

[12] C. Visweswariah, *et.al.* First order incremental block based statistical timing analysis. In *Proc.DAC*, pages 331–336, 2004.

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

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

[15] F. Liu. A general framework for spatial correlation modeling in vlsi desing. In *Proc. DAC*, pages 817–822, 2007.

[16] W. Meeker and L. Escobar. *Statistical methods for reliability data.* Wiley, 1998.

[17] S. Kotz, *et.al. Continuous Multivariate Distributions.* Wiley, 2000.

[18] Q. Liu, *et.al.* Confidence scalable post-silicon statistical delay prediction under process variation. In *Proc. DAC*, pages 496–502, 2007.

[19] S. Reda, *et.al.* Analyzing the impact of process variations on parametric measurements: Novel models and applications. In *Proc. DATE*, pages 375–380, 2009.

[20] N. Cressie. *Statistics for spatial data.* Wiley, 1993.

[21] K.-H. Yuan and P. M. Bentler. Two simple approximations to the distributions of quadratic forms. Paper 2007010106, Department of Statistics, UCLA, 2007.