Process Variation and Temperature-Aware Reliability Management

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

Abstract*—In aggressively scaled technologies, reliability concerns such as oxide breakdown have become a key issue. Dynamic reliability management (DRM) has been proposed as a mechanism to dynamically explore the trade-off between system performance and reliability margin. However, existing DRM methods are hampered by the fact that they do not accurately model spatial and temporal variations in process and temperature parameters which have a strong impact on chip reliability. In addition, they make the simplifying assumption that the future workloads are identical to the currently observed one. This makes them sensitive to sudden workload variations and outliers. In this paper, we present a novel workload-aware dynamic reliability management framework that accounts for local variations in both the process and temperature. The reliability estimation, along with the predicted remaining workload is fed to a dynamic voltage/frequency scaling module to manage the system reliability and optimize processor performance. Using a fast on-line analytical/table-look-up method we demonstrate an average error of 1% with up to 5 orders of magnitude speedup compared to Monte Carlo simulation. Experiments on an Alpha-like processor show our DRM framework fully utilizes the available margin and achieves 28.7% performance improvement on average.

I. INTRODUCTION

Increasing manufacturing and environmental variabilities create new challenges in designing a reliable system in nano-scale technologies [1]. Traditionally, chips are designed under the assumption of worst-case temperature and supply voltage conditions to ensure a certain reliability criterion throughout the lifetime [2]. Such guard-band approach approximates the design as a uniform static transistor network and ignores the dynamic nature of environmental conditions (temperature, supply voltage, workload, *etc.*) and the static non-uniform nature of process parameter variations. As a result, the approach is overly conservative and incurs significant area/performance/power cost.

To resolve these issues, several dynamic mechanisms such as dynamic reliability management (DRM) and dynamic temperature management (DTM) have been proposed to recapture the system performance and reliability margin [2]–[6]. Unlike DTM, which tends to avoid temperature violations to minimize the chances of reliability failure, DRM reacts to the current reliability state and then adjusts the voltage/frequency to recover unacceptable reliability loss or use available excess margin to boost performance by allowing operation at a higher supply voltage [3]–[5]. Thus, DRM enables valid assessment and control of system reliability for a user-defined lifetime target.

The concept of DRM was first introduced in [5], which considered multiple failure mechanisms. Reference [6] extended the approach to monitor and control the impact on lifetime reliability, by either thread scheduling or dynamic voltage/frequency scaling (DVFS). In the latter case, the current level of reliability degradation is used to set a voltage limit that constrains the DVFS algorithm. The chip lifetime is therefore guaranteed to be met with a high confidence level while maximizing the allowable performance [6]. Karl, *et.al.*, also explored the DRM framework using a systematic model to improve the performance [2]. However, since the existing DRM approaches only partially consider the manufacturing/environmental variabilities, they can incur inaccuracies or unnecessary margins in realistic reliability assessment:

- First, none of [2], [5], [6] incorporate process variation in the reliability prediction. Instead they rely on deterministic worst-case process parameter values. As process variation has become ever more pressing, such deterministic analysis is inaccurate and degrades the efficiency of DRM.
- Second, many existing approaches only capture the temporal temperature variation but neglect its spatial variation. Most permanent wear-out mechanisms [7], such as oxide breakdown (OBD), electromigration (EM) or thermal cycling, have an exponential dependence on temperature. Thus, neglect of temperature difference among functional blocks inevitably induces inaccuracy in results.
- Finally, existing DRM frameworks predict reliability by assuming that the future workload is identical to the current one. However, this assumption is very sensitive to sudden workload variations. If the workload suddenly spikes, the DRM will erroneously extrapolate this workload for the remainder of the lifetime and limit the supply voltage overly strictly, thereby sacrificing possible performance gain. On the other hand, if the workload suddenly drops, the assumption that it will remain low throughout the remainder of the chip lifetime will set the voltage limit too loosely, thereby allowing a sudden high workload to consume most of the remaining reliability budget. If this occurs near the end of the lifetime, a final high workload could cause a chip failure before the required life time is met, which would constitute a DRM failure.

In this paper, we propose a novel workload-aware dynamic reliability management framework accommodating both the process and temperature variations into the lifetime reliability analysis. Among all permanent failure mechanisms, OBD has become increasingly prominent due to the aggressive thickness scaling and hence is the initial focus of our paper. After that we outline how to incorporate other permanent failure mechanisms in our DRM framework. The proposed framework has the following key modeling contributions:

- Process variation-aware OBD reliability assessment. Prior OBD analysis either assumes oxide thickness uniformity across the chip [8] or discards the temperature variations [9] and hence cannot be employed by dynamic systems. To resolve this problem we extend the approach in [8], [9] and project the tremendous parameter space of device-level oxide thicknesses to the architecture-block level by characterizing the block-level oxide thickness distribution. The system reliability is then compactly expressed as the sum of N double integrals for the N blocks and can be efficiently computed.
- Handling of spatial/temporal temperature variations. The evaluation of OBD reliability at the granularity of blocks naturally enables the incorporation of block-level spatial temperature variation into the analysis [1]. To handle the temporal variation, we partition the lifetime into a series of time frames to investigate how OBD reliability evolves with time and retains the cumulative damage of its past states. Moreover, we also formulate the proposed process variation and temperature-aware OBD model with a hybrid analytical/table-look-up approach, which enables extremely fast on-line computation. These key features (low complexity and variation awareness) are crucial to effective reliability assessment of a dynamic system.
- Confidence-based workload prediction. The DRM framework in this paper uses a confidence-based workload prediction scheme to anticipate the statistical characteristics of the long-term workload. We examine the long-term workload temporal correlation and use it to predict the mean of the remaining workload subject to a certain level of confidence. The impact of future workloads is then evaluated to improve the decisions on reliability control.
- DVFS control. We finally show how all these components can be combined to formulate a comprehensive DRM framework using a DVFS controlling scheme. A design flow to integrate a proportional-integral-derivative (PID) controller into the proposed nonlinear DRM system is introduced, which facilitates voltage/frequency adjustment in the DVFS module.

We employ several benchmarks to verify the efficacy of our DRM framework and underlying OBD reliability model. Experimental results show that with the hybrid analytical/table-look-up implementation, our reliabilty model can achieve an average error of 1% in lifetime estimation for six different designs as well as five orders of magnitude speedup compared with Monte Carlo simulation. Based on the model, the framework can accurately control an Alpha processor to meet the predefined lifetime with average performance improvement of 28.7%.

[1] This approach assumes temperature to be uniform across a block. However, large blocks can be partitioned into smaller block sizes if this assumption does not hold.

978-3-9810801-6-2/DATE10 © 2010 EDAA

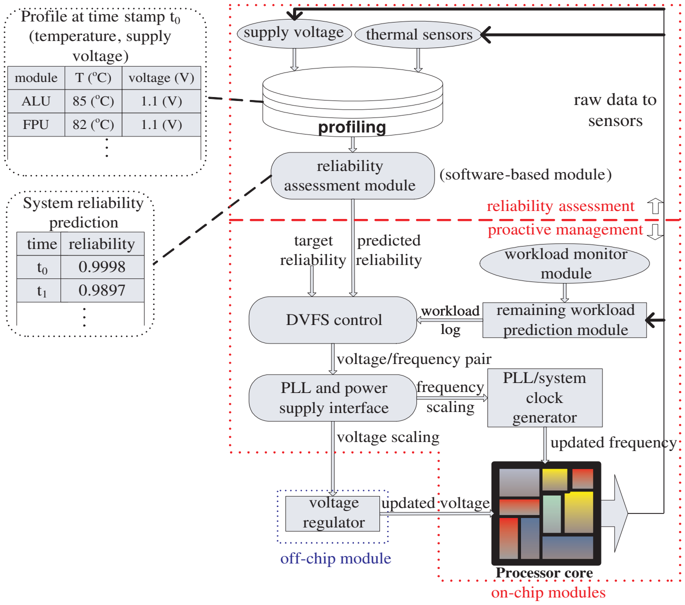


Fig. 1. A high-level block diagram of the proposed DRM system

II. Design Flow of the Proposed DRM System

The proposed process variation and temperature-aware DRM framework is shown in Figure 1. The framework performs two main tasks, reliability assessment and proactive management. The first task, on-line reliability evaluation and prediction (top half of Figure 1), is achieved by the reliability assessment module. This module is software-based and takes into account both process and temperature variabilities for system reliability monitoring, as detailed in section III.

The input to the assessment module is the system profile including temperature and supply voltage information of different functional blocks. The raw data streams for the profile can be collected from the circuit-level sensors. For design-stage simulation purpose, we collect the thermal information by using Wattch [10] to monitor the power consumption of functional blocks and then feeding to HotSpot [11] to achieve the temperature profiles of the chip. The on-line profiling module then processes the raw data and generates a comprehensive architecture-block-level system profile, a typical case of which is shown on the top left of Figure 1. The time-varying system profile helps determine the parameters of the reliability models for the chip and is used to capture the spatial and temporal temperature/voltage variations. The hybrid analytical/table-look-up method used in our assessment module can easily incorporate the variation information into analysis, enabling efficient on-line prediction.

The second task of the proposed DRM framework is proactive management of reliability (bottom half of Figure 1), and is achieved by the DVFS module. This module computes the deviation between the predicted overall lifetime reliability and target reliability. This deviation, along with the predicted remaining workload, is then used in the control system (PID) of the DVFS to adjust the system maximum voltage for the next time-stamp. Unlike general-purpose dynamic management systems, DRM is a management strategy targeted at long-term system reliability control and monitors the workload at a macro-level of processor usage instead of a instruction-set level [2]. The monitored workload is used to predict the remaining workload. Due to the uniqueness of DRM, our emphasis is placed upon the long-term temporal characteristics of workloads. By noting that most existing workload prediction researches focus on phase-scale characteristics and cannot be directly employed by DRM, we introduce a confidence-based workload prediction method to exploit its long-term temporal behavior, as described in section IV.

The maximum voltage set by the PID controller limits the possible voltage that the system can use during high workload conditions. In the following time, the operating voltage/frequency chosen by the DVFS (limited by the computed maximum voltage) is fed into the processor core (through the PLL/supply voltage interface, clock signal generator and voltage regulator). Since DRM system does not have critical demands on response time, either workload prediction module or DVFS module can be implemented in software. The design details of the DVFS module is discussed in section V. Once the supply voltage/clock frequency for the chip is updated, the raw data of temperature/voltage information is again collected by the sensors and sent to the profiling module to complete the control loop.

III. Process Variation and Temperature-Aware Full-Chip Reliability Analysis

This section discusses the statistical model and implementation of the reliability assessment module in Figure 1. As stated in section I, this paper focuses on the process variation and temperature-aware model for the OBD failure, while adopting the existing models in [2], [7] for the other permanent wear-out mechanisms to comprehensively evaluate the system reliability. Given a chip with N blocks and m devices, we define the following notations in Table I for the remainder of the paper.

TABLE I
Notations used in the full-chip reliability analysis

Notation	Definition
$\mathbf{x} = [x_1, \ldots, x_m]$	oxide thicknesses for m device of a chip
a_i	area for the i_{th} device of the chip, $i = 1...m$
A_j	total area for the j_{th} block of the chip, $j = 1...N$
m_j	the number of devices within the j_{th} block
$f_{\mathbf{x},\mathbf{y}}(x, y)$	joint probability density function (PDF) of x and y, where x and y can be either vector or scalar

A. Review of Full-Chip Oxide Breakdown Reliability Analysis

The gate oxide degradation is a non-deterministic process dependent on the oxide-thickness, transistor area, voltage, and temperature [12]. The oxide breakdown time is typically modeled as a random variable (RV) following a Weibull probability distribution function [13]:

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

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 b are the scale and shape parameters of the Weibull distribution, and x is the oxide thickness of the device [13]. Both parameters α and b depend on temperature and voltage and can be expressed by the closed-form models[2] [13]–[15]. Then, for a given oxide thickness, the reliability function of a device can be interpreted as its conditional reliability function and written as $R(t|x) = e^{-a(\frac{t}{\alpha})^{bx}}$. The overall reliability function for the whole chip is given by:

$$R_c(t) = \int_0^{\infty} \ldots \int_0^{\infty} \prod_{i=1}^{m} R_i(t|x_i) f(x_1 \ldots x_m) dx_1 \ldots dx_m \tag{2}$$

where $f(x_1, \ldots, x_m)$ is the joint probability density function (PDF) of the gate oxide thicknesses for m devices.

To simplify the product $\prod_{i=1}^{m} R_i(t|x_i)$, [9] proposed to use a conditional probability $R_c(t|u, v)$ where u and v are the sample mean and variance of the chip-level oxide thickness distribution. Thus, (2) is compactly expressed by:

$$R_c(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_c(t|u, v) f_{\mathbf{uv}}(u, v) du\, dv \tag{3}$$

where

$$R_c(t|u, v) = \exp(-Ae^{\ln(\frac{t}{\alpha})bu + (\ln(\frac{t}{\alpha}))^2 b^2 v/2}) \tag{4}$$

and $f_{\mathbf{uv}}(u, v)$ is the joint PDF of u and v.

B. Incorporating Block-Level Spatial Temperature Variation

A potential problem of (3) is the underlying assumption that all the devices across the chip share the same worst-case temperature/voltage and hence bear the same parameters b and α for the device reliability function $R(t|x_i)$. However, in reality, the on-chip temperature may vary from block to block as shown in Figure 2. It is known that both parameters b and α are heavily dependent on temperature [14], [15]. Thus, it is unfair to assume that hot-spots and inactive areas have the same reliability model and are equally prone to the OBD failure. To

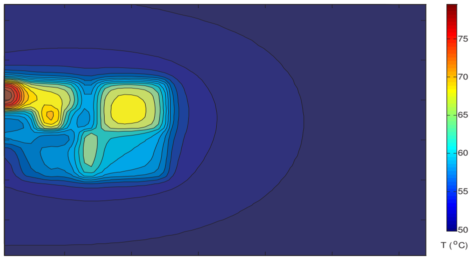


Fig. 2. Temperature profile for an Alpha processor simulated by HotSpot [11]

overcome this problem, we need to account for temperature variations in the reliability model.

In practice, temperature varies continuously across the chip. Transistors within a particular architecture-level block may share similar temperature due to the similar activities and supply voltage. On the contrary, inter-block temperature variation is much higher as functional blocks usually perform completely different operations [3]. It is therefore sufficient to construct a temperature-aware reliability model at the granularity of architecture-level blocks.

Then, given the oxide thickness, a theoretically rigorous formulation of full-chip OBD reliability model can be described as the following:

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i) = e^{-\sum_{i=1}^{m} a_i (\frac{t}{\alpha_i})^{b_i x_i}} \tag{5}$$

where α_i and b_i indicate temperature-dependent parameters for the device-level reliability function in $R(t|x)$. Because of the negligible temperature difference within one block, (5) can be expressed as:

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i) = e^{-\sum_{j=1}^{N} \sum_{i=1}^{m_j} a_{i,j} (\frac{t}{\alpha_j})^{b_j x_{i,j}}} \tag{6}$$

where N is the number of architecture-level blocks.

Typically, the parameter variation of a device includes inter-chip, intra-chip spatially correlated and random variation components [16]. Thus, for a set of devices within one particular block, they may have different oxide thicknesses due to the variability. The frequency distribution histogram of the oxide thicknesses for this block can then be interpreted as a Block-level Oxide-thickness Distribution (**BOD**). BOD shows how many devices in this block correspond to a particular oxide thickness. Due to the high spatial correlation within one block, all the devices within the block have the same inter-chip variation component and approximately the same spatially-correlated variation component. Thus, the difference in oxide thickness within the block is mainly caused by random variation. BOD can therefore be considered as the histogram of oxide thickness samples for different devices that are independently drawn from one Gaussian random variable. As the number of devices increases, **the histogram of the oxide thickness samples (or BOD) follows the curve of a Gaussian distribution.** Figure 3 validates the claim with the oxide thickness histograms for two blocks (a decoder and a 32-bit multiplier) from a design similar to Alpha 21264. It is apparent that we get distinctly Gaussian-like curves.

Now, by discretizing the BOD for a block into k_j discrete intervals assuming a truncated distribution, (6) can be represented by:

$$R_c(t|\mathbf{x}) = e^{-\sum_{j=1}^{N} \sum_{i=1}^{k_j} \overline{a}_{i,j} (\frac{t}{\alpha_j})^{b_j \overline{x}_{i,j}}} \tag{7}$$

where $\overline{x}_{i,j}$ denotes the midpoint of the i^{th} discrete interval of BOD for the j_{th} block and $\overline{a}_{i,j}$ is the total area of all devices having thickness $\overline{x}_{i,j}$. Thus, after normalizing the exponent with the total area of each functional block, (7) can be expressed as:

$$\begin{aligned} R_c(t|\mathbf{x}) = R_c(t|\mathbf{u},\mathbf{v}) &= \exp\left[-\sum_{j=1}^{N} A_j \sum_{i=1}^{k_j} p_{i,j} (\frac{t}{\alpha_j})^{b_j \overline{x}_{i,j}}\right] \\ &\approx \prod_{j=1}^{N} \exp[-A_j \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha_j})^{b_j x} dx] \end{aligned} \tag{8}$$

where $p_{i,j} = \overline{a}_{i,j}/A_j$ is the probability of observing an oxide-thickness $\overline{x}_{i,j}$ within a particular block; $\phi(x)$ is the PDF of a standard Gaussian; u_j (v_j) is the mean (variance) of the j_{th} BOD.

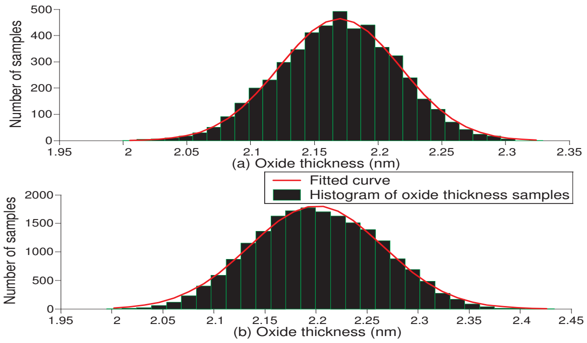


Fig. 3. Histograms of the oxide thicknesses for (a) a decoder and (b) a multiplier

Based on (8), the conditional reliability $R_c(t|\mathbf{x})$ is compactly expressed by $R_c(t|\mathbf{u},\mathbf{v})$ with $2N$ distinct variables, where $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ are the corresponding mean and variance of BODs for N blocks. For the lifetime period of interest, (8) can be further simplified using Taylor expansion:

$$R_c(t|\mathbf{u},\mathbf{v}) \approx 1 - \sum_{j=1}^{N} \left[1 - e^{-A_j \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha_j})^{b_j x} dx}\right] \tag{9}$$

where the integral in the exponent is:

$$g(u_j, v_j) = e^{\ln(\frac{t}{\alpha_j}) b_j u_j + (\ln(\frac{t}{\alpha_j}))^2 b_j^2 v_j/2} \tag{10}$$

$g(u_j, v_j)$ is used for clarity throughout the rest of the paper.

Then we can integrate the conditional reliability function in (9) over the joint PDF $f_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v})$ to evaluate the overall reliability:

$$R_c(t) = \oint \oint \left[1 - \sum_{j=1}^{N} \left(1 - e^{-A_j g(u_j,v_j)}\right)\right] f_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) d\mathbf{u} d\mathbf{v} \tag{11}$$

Since $\exp[-A_j g(u_j, v_j)]$ is independent of any u_i or v_i that $i \neq j$, the equation above is variable separable and can be compactly represented by the summation of N double integrals:

$$R_c(t) = 1 - N + \sum_{j=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-A_j g(u_j,v_j)} f_{\mathbf{u_j},\mathbf{v_j}}(u_j, v_j) du_j dv_j \tag{12}$$

For each double integral in (12), we employ the marginalization technique in [9] to represent the joint PDF $f_{\mathbf{u_j},\mathbf{v_j}}(u_j, v_j)$ as the marginal PDF product, where $f_{\mathbf{u_j}}(u_j)$ is a Gaussian PDF and $f_{\mathbf{v_j}}(v_j)$ is a chi-square PDF for the j_{th} BOD.

C. From Static to Dynamic Model

The spatial temperature variation-aware model in section III-B is still a static model assuming the same temperature/voltage profiles over the entire time-span, which cannot be applied to DRM. In practice, for a dynamic system, both temperature and supply voltage are varying from time to time. In this section we will investigate how to extend the proposed model for a dynamic system.

For any time period, we can divide it into k time frames, [0,Δt], [Δt,2Δt]... [$(k-1)\Delta t$, $k\Delta t$]. When Δt is small, it is reasonable to assume that the supply voltage and temperature profile are steady within the frame. The dynamic system reliability $R_d(k\Delta t)$ at any time $k\Delta t$ is then the probability that the system does not fail within any single frame in the period [0, $k\Delta t$]:

$$R_d(k\Delta t) = \prod_{i=1}^{k} [1 - (R_i((i-1)\Delta t) - R_i(i\Delta t))] \tag{13}$$

where $R_i(t)$ is the static reliability function of (12) within the time frame [$(i-1)\Delta t$, $i\Delta t$]. Since supply voltage and temperature profile are steady within this time frame, the shape/scale parameters of $R(t|x_i)$ are then uniquely determined and lead to a steady $R_i(t)$.

$R_i((i-1)\Delta t) - R_i(i\Delta t)$ can be interpreted as the damage to the system within the i_{th} time frame and is small in practice. With Taylor expansion, we can rewrite (13) as:

$$R_d(k\Delta t) \approx 1 - \sum_{i=1}^{k} [R_i((i-1)\Delta t) - R_i(i\Delta t)] \tag{14}$$

[2] Our work employs linear models for $\ln(\alpha)$ and b as in [13]–[15], but is not limited by the particular forms of models.

where the summation part denotes the accumulated damage during the past k time frames. This on-line model for dynamic system reliability estimation incorporates both temporal and spatial temperature variations and can be incrementally evaluated, *i.e.*, only the reliability function within the current time frame needs computation.

D. Extended to a Fast Hybrid Analytical/Table-Look-Up Method

In this section we combine the analytical model in (12) with a table look-up method to achieve further speed up. The pre-calculated look-up table only needs to be computed once for a particular chip and can be used throughout the lifetime.

Take the j_{th} integral in (12) for example. Since u_j and v_j are integration variables and eliminated afterwards, the result of the integral is fully determined by A_j and the parameters of $g(u_j, v_j)$ ($\ln(t/\alpha_j)$ and b_j as in (10)). Once the chip is designed, A_j happens to be a constant for the j_{th} functional block. Thus, with $\ln(t/\alpha)$ and b acting as indices we can construct a two-dimension look-up table to compute the double integral for each functional block[3]. Then, the system reliability at any time t under certain temperature/voltage conditions (which determine the values of α and b), can be easily computed using bilinear interpolation according to the indices of $\ln(t/\alpha)$ and b. For N functional blocks, we have N look-up tables, with $n_\alpha \times n_b$ entries in each table, where n_α(=100) and n_b(=100) are the number of indices for parameters $\ln(t/\alpha)$ and b, respectively. Experimental results in section VI show that the hybrid method leads to a faster speed but nearly equivalent accuracy when compared to the analytical model in (12).

E. System Reliability for Multiple Permanent Failure Mechanisms

Similar as [2], the total chip reliability is derived by using the contributions of each failure mechanism (OBD, EM, thermal cycling, *etc.*). Due to the space limitation, we here only outline the computation flow while the model details can be found in [2], [7]:

- *Compute the failure probability of different mechanisms.* The OBD model is discussed in details in sections III-B and III-C. For any other mechanism, we first derive its MTTF and then use the Miner's rule and a Weibull distribution model to convert the percentage of lifetime to failure probability [2], [7].
- *Evaluate system reliability.* Once the failure rate/reliability of all considered mechanisms are achieved, the system reliability can be comprehensively evaluated as in [2]:

$$R_{sys} = (1 - P_{OBD})(1 - P_{EM})(1 - P_{TC}) \tag{15}$$

where P_{OBD}, P_{EM} and P_{TC} are the failure rate due to OBD, EM and thermal cycling, respectively. Since in our model temperaure/voltage profile is known and directly used as the primary input for calculation, this strategy naturally captures the correlation and enables efficient on-line control of the system reliability.

IV. Confidence-Based Workload Prediction

DRM is a management strategy focusing on long-term system behavior. Existing DRM works [2], [6] perform prediction based on the assumption that the future workload remains the same as the current one. Such prediction may be easily disturbed by a short impulse and thus make overly aggressive decisions. It is therefore crucial to fully explore the long-term temporal correlation of the workloads and predict their statistical characteristics.

We define $W[n]$ as the normalized workload monitored at time stamp nT, where T is the monitoring interval between two stamps. Figure 4 illustrates a 24-hour workload trace collected from an actual server. There are two observations about the figure: (1) effect of locality, *i.e.*, similar workloads are likely to be carried out in the near future; (2) low or even zero long-term temporal correlation, mainly because the threads executed at some time stamp can be completed within a certain period of kT, where kT is the maximum execution time of a thread.

Based on those observations, we divide the executed workload trace into M chunks, where $W[(i-1)p+1], ..., W[ip]$, are p monitored workload values in the i_{th} chunk. Each chunk is then associated with a random variable (RV), X_i, where $i = 1, 2, ...M$ and M increases with time. The p workload values within the i_{th} chunk can be interpreted as the samples of X_i. Figure 5(a) shows the correlation contour between X_i's of the workload trace in Figure 4, where the diagonal line denotes the self-correlation coefficient of 1, and any off-diagonal entry at (i, j) denotes the cross-correlation coefficient between X_i and X_j. Clearly, most cross-correlation coefficients are zero and in agreement with the claim of un-correlation for long-term scale workloads. Although in theory un-correlation does not guarantee the independence of non-Gaussian RVs, we find that assuming independence for X_is is a reasonable approximation in practice. We then denote the sample mean as $\overline{X} = \frac{X_1+X_2+...+X_M}{M}$, which approaches a **Gaussian RV** as M increases. Thus, $E(\overline{X})$ turns out to be an estimator of the mean of the entire workload trace, *i.e.*, $E(\overline{X}) = E(W)$.

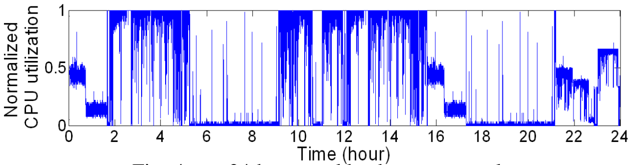
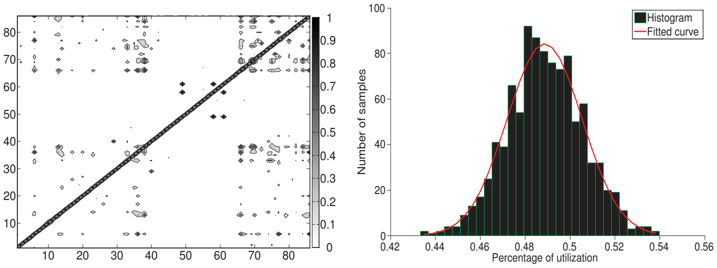


Fig. 4. a 24-hour workload trace example



(a) (b)

Fig. 5. (a) Correlation contour of the workload trace (b) Histogram of $\overline{X}$ and the fitted gaussian curve for the workload trace in Figure 4 with M=86 and p=1000 and T=1 second

According to the central limit theorem, the distribution of $\overline{X}$ can be constructed by averaging the randomly-selected samples from X_i's (M chunks). Figure 5(b) displays the histogram of $\overline{X}$ for the workload trace in Figure 4, which is very close to a Gaussian curve. Once $E(\overline{X})$ is achieved from the histogram, we can estimate the mean of the remaining workload, $E_{remain}(W)$ by:

$$E_{remain}(W) = \frac{E(\overline{X}) \times T_{life} - E_{exe}(W) \times T_{current}}{T_{life} - T_{current}} \tag{16}$$

where T_{life} is the pre-defined lifetime, $T_{current}$ is the current time stamp and $E_{exe}(W)$ is the mean of the executed workload.

A too conservative estimation may limit voltage scaling and hence reduce the potential performance improvement, whereas an overly aggressive prediction increases the chances of wear-out before the specified minimum operational lifetime. From the standpoint of reliability, an aggressive prediction should be avoided in the period when prior knowledge about the workload is not available or the accumulated damage is close to the pre-defined threshold. We therefore employ a confidence-based scheme to ensure that $E(W)$ is not over-aggressively predicted. Then $E(\overline{X})$ in (16) is replaced by a confidence-based estimator X_{esti}, the confidence of which is defined as $C = Pr(\overline{X} \leq X_{esti})$ to evaluate the degree of conservativeness of the prediction. In other words, once the confidence of the prediction is determined, we can compute X_{esti} according to the distribution of $\overline{X}$.

Clearly, the system needs to be conservative at the early stage or the end, and aggressive in the middle of the lifetime. Based on such intuitions, we extend a bathtub-like failure function in [17] to two-dimensions in Figure 6. The x-axis and y-axis denote the normalized failure rate $F_{norm} = (1 - R_{sys})/(1 - R_{target})$ and normalized lifetime $T_{norm} = T_{current}/T_{life}$, respectively, where R_{sys} is the actual chip reliability and R_{target} is the reliability target within the predefined lifetime. The confidence is then given by:

$$C = \begin{cases} \lambda\beta(x_0 x)^{\beta-1} e^{(x_0 x)^\beta} + C_0, & x \leq 0.5 \\ \lambda\beta(x_0(1-x))^{\beta-1} e^{(x_0(1-x))^\beta} + C_0, & x > 0.5 \end{cases} \tag{17}$$

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

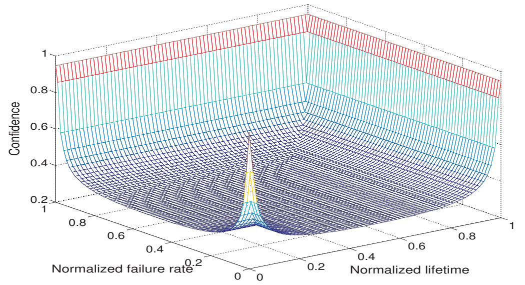


Fig. 6. Bathtub-like confidence function with respect to normalized lifetime and normalized failure rate

where λ, β, x_0 and C_0 are parameters from [17] and $x = max(F_{norm}, T_{norm})$ [4]. The complete algorithm for on-line confidence-based workload prediction is summarized as follows:

Step 1: Compute the confidence C using (17);
Step 2: Compute p samples for $\overline{X}$ by averaging the randomly-selected samples from X_i's (M chunks);
Step 3: Construct the histogram for $\overline{X}$;
Step 4: Fit the histogram to a Gaussian distribution;
Step 5: Compute X_{esti} that satisfies $C = Pr(\overline{X} \leq X_{esti})$;
Step 6: Compute $E_{remain}(W)$ using (16);

V. Design Flow of DVFS Module

Due to the nonlinearity of the statistical reliability model, we suggest a two-stage design flow of PID controller for DVFS. First, we replace the nonlinear system with a simplified linear system and derive the characteristic function. The stability condition along with the performance constraints define the feasible domain of the PID parameters. Second, we restore the original nonlinear system and fine-tune the parameters within the feasible domain to maximize the performance gain.

We approximate the system by using a deterministic model for linearization. The system reliability is then:

$$R_{sys}(t) = R_0 \exp(-A(t/\alpha)^{(bx)}) \qquad (18)$$

where R_0 conservatively bounds the impacts of EM and thermal cycling. For a certain reliability target $R_{sys,target}$, system lifetime can be formulated by:

$$\ln(t) = \ln(\alpha) + \ln((-\ln(R_{sys,target}/R_0)/A)^{(1/bx)}) \qquad (19)$$

The target and predicted reliability are then compared in the form of the log of lifetime (reference signal $\ln(t_{target})$ versus $\ln(t_{predict})$). Since $\ln(\alpha)$ is linearly dependent on supply voltage, the sub-system described in (19) is hence a linear system. By noting the discrete nature of the system, the z-domain model can be given by $G_s(z) = K_\alpha/z$ [18].

Assume the z-domain model for the PID controller is: $G_p(z) = K_p + K_I\frac{z}{z-1} + K_D\frac{z-1}{z}$. The characteristic function is then a three-order function of z that $T(z) = z^3 + K_1z^2 + K_2z + K_3$, where K_1=K_α(K_I+K_D+K_p)-1, K_2=-K_α(2K_D+K_p) and K_3=$K_\alpha K_D$ [18]. By Vieta's Formulas, the relationship between the roots of $T(z)$ and PID parameters can be explored. Instead of selecting the roots within the unit circle to determine the parameters of PID [18], we utilize the equations above as well as the constraints on predefined settling/rise time and maximum overshoot to obtain the feasible domain. Similar as [2], the feasible domain is then swept to find the parameter pair achieving the maximum performance for the actual nonlinear system.

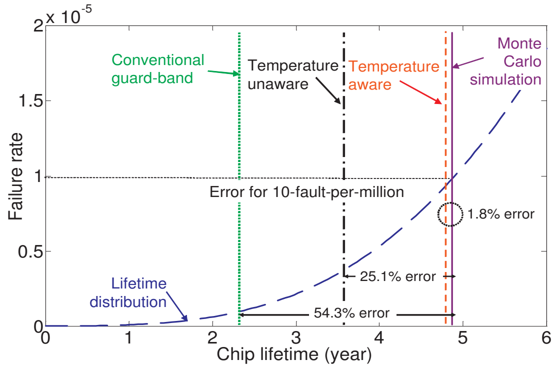


Fig. 7. Errors of the 10-faults-per-million for Monte Carlo simulation, the proposed tempearture-aware approach, temperature-unaware approach in [9] and conventional guard-band assuming minimum oxide thickness

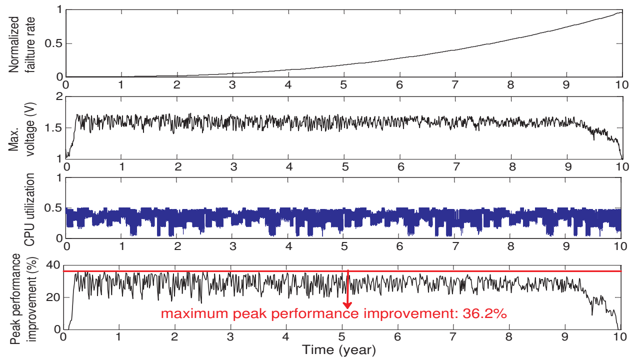


Fig. 8. DRM system behavior for a 10-year workload trace

VI. Experimental Results

The proposed variation-aware reliability model and DRM system were implemented and tested on several benchmarks (C1-C6) varying from 50K to 0.84M devices. All the process parameters for the designs are based on a commercial 130nm technology model. The oxide thickness variation model includes the inter-die, intra-die spatially correlated and random variation components [16]. The defect generation relationships for the technology node and the technology dependent parameters of the oxide reliability function model are obtained from [13]–[15]. The workload traces in the experiments were collected from several servers for several months to provide realistic information about processor utilization. The temperature profile of the system is simulated by HotSpot [11] with Wattch to estimate the block power [10].

A. Efficacy of the Proposed Variation-Aware OBD Model

Once the post-layout design implementation and temperature profile are given, we use the method in section III-B and III-D to compute the OBD reliability. To validate the results, we employ the Monte Carlo (MC) simulation of 1000 samples to compute the OBD reliability. Table II compares our temperature-aware statistical approach (statistical) in section III-B and the hybrid analytical/table-look-up approach (hybrid) in section III-D with MC simulation. Columns 3-6 compare the accuracy based on lifetime estimation for 1-fault-per-million parts and 10-faults-per-million parts. It is clear that both methods are in good agreement with MC simulation, with errors of around 1% on average. Columns 7-11 compare the run time for three methods. Unlike MC simulation, both our statistical approach and hybrid approach are able to analyze the circuit in seconds, independent of the number of devices. The statistical approach (statistical) demonstrates around 2-3 orders of magnitude speedup for all the designs, whereas MC simulation scales super-linearly with the number of devices. Moreover, the hybrid approach (hybrid) shows more than five orders of magnitude speedup over the MC simulation.

We further compare the overall reliability estimation results in Figure 7 for design C3 using MC simulation, the proposed approach in section III-B, temperature-unaware approach from [9] and conventional guard-band approach that assumes minimum oxide thicknesses across the chip. The chip lifetime distribution (blue curve) is achieved by MC simulation with 10000 sample chips of C3. One can see that for 10-faults-per-million criterion, the temperature-unaware approach and the guard-band have 25.1% and 54.3% errors, whereas our temperature-aware approach can achieve an accuracy of 1.8% error and is very close to the result by MC simulation. This clearly indicates the necessity for a process variation and temperature-aware approach for reliability analysis.

B. Performance of DRM System

We implemented the proposed DRM on a five-stage pipeline processor similar to Alpha 21264 with 15 functional blocks. Each block is mapped to the device-level for the reliability model in section III. The voltage range for DVFS is 0.8-1.8V with a nominal voltage of 1.2V.

[4] This function is not unique and can be replaced by other similar formulations.

TABLE II
ACCURACY AND RUN TIME COMPARISON OF THE TEMPERATURE-AWARE STATISTICAL APPROACH IN SECTION III-B AND HYBRID ANALYTICAL/TABLE-LOOK-UP APPROACH IN SECTION III-D WITH MONTE CARLO SIMULATION

Design	No. of devices	Lifetime estimation error (%) w.r.t. MC				Run time (sec.)/Speed up w.r.t. MC				
		1/million		10/million		statistical		hybrid		MC
		statistical	hybrid	statistical	hybrid					
C1	50K	0.84	0.12	1.18	1.84	1.51	177×	0.020	13498×	267
C2	80K	1.50	0.68	1.28	0.30	1.59	238×	0.022	17486×	380
C3	0.1M	2.04	0.16	1.77	2.26	1.92	245×	0.019	24122×	470
C4	0.2M	2.23	0.63	1.90	1.30	1.93	363×	0.020	35206×	702
C5	0.5M	0.20	3.42	0.12	1.65	1.86	837×	0.020	77845×	1557
C6	0.84M	0.64	1.63	0.54	0.76	1.95	1183×	0.020	115325×	2307
Average		1.24	1.11	1.13	1.35		418×		47247×	

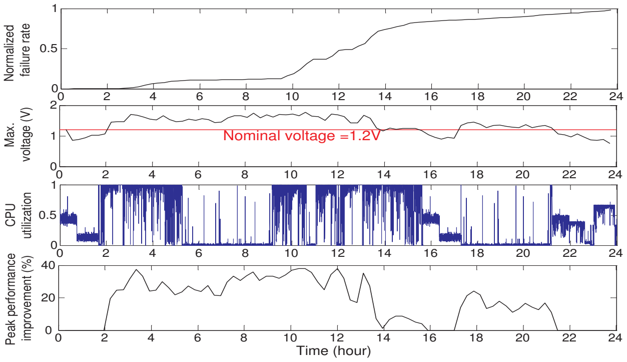


Fig. 9. DRM system behavior for a 24-hour workload trace

Figure 8 shows DRM system response for a 10-year period with a randomized combination of workload traces collected from the servers. The voltage is updated by PID controller on an hourly basis with workload monitored on a second basis. The 1^{st} sub-figure shows the normalized failure rate of the chip, which is close to the target at the end. The maximum voltage in the 2^{nd} sub-figure represents the voltage limit allowed for DVFS, not the actual operating voltage. The 3^{rd} sub-figure is the normalized CPU utilization (workload). The 4^{th} sub-figure shows the peak performance improvement as a measure of the improvement in attainable frequency (%) during periods of peak CPU demand. The system demonstrates steady behavior for this trace and achieves 28.7% gain on average and 36.2% gain at maximum.

We also explore a short-scale trace in Figure 9 to illustrate the details of the DRM system, especially the interaction between the normalized failure rate, maximum supply voltage, workload and peak performance improvement. Figure 9 presents the DRM simulation for a 24-hour workload trace, with voltage updated every ten minutes. It is apparent that the degradation becomes more aggressive in the middle stage of the trace, whereas it maintains a slow pace at the beginning and the end. That is because the system tends to be conservative at the early stage or the end in case of unpredicted wear-out, and hence assigns a lower maximum voltage. The last sub-figure in Figure 9 shows peak performance gain of the proposed DRM, which can achieve 17.5% improvement on average and 37.6% at maximum.

Figure 10 displays the workload prediction results for the 10-year workload trace, including the actual mean values of the whole trace $E(W)$ and the remaining workload, the confidence-based workload estimation X_{esti}, and its mean estimation $E_{remain}(W)$. One can see the conservativeness of the prediction at the beginning and the end and how the estimation values track the actual values in the middle with certain aggressiveness. Figure 11 illustrates the reaction of the DRM system to a sudden workload change at the 5^{th} year, demonstrating the system ability to control sudden changes.

VII. CONCLUSIONS

This paper presents a workload-aware dynamic reliability management framework based on a statistical OBD reliability model. The underlying OBD model captures the process and temperature variations and accurately estimates the system reliability. A confidence-based workload prediction scheme further improves the performance gain that can be achieved. Experimental results show that the on-chip real-time reliability monitoring and control can boost supply voltages beyond nominal values, which may enhance system performance by 28.7% without violating pre-defined reliability constraints.

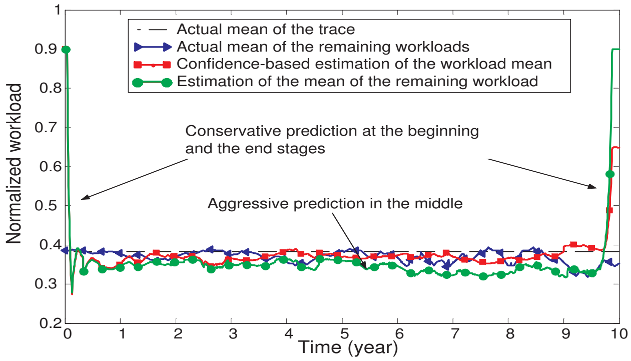


Fig. 10. Confidence-based workload prediction for a 10-year workload trace

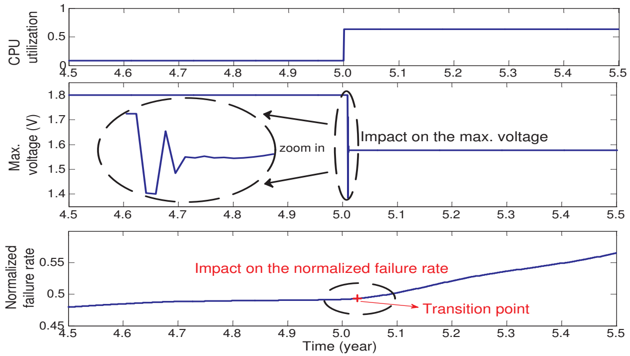


Fig. 11. DRM response to a sudden change on workload

VIII. ACKNOWLEDGEMENT

The authors gratefully acknowledge the Semiconductor Research Corporation and National Science Foundation for supporting this work.

REFERENCES

[1] S. Borkar. Designing reliable systems from unreliable components: the challenges of transistor variability and degradation. *IEEE Micro.*, 25(6):10–16, 2005.
[2] E. Karl, *et.al.* Reliability modeling and management in dynamic microprocessor-based systems. In *Proc. DAC*, pp. 1057–1060, 2006.
[3] K. Skadron, *et.al.* Temperature-aware microarchitecture. In *ISCA*, pp. 2–13, 2003.
[4] Z. Lu, *et.al.* Interconnect lifetime prediction under dynamic stress for reliability-aware design. In *Proc. ICCAD*, pp. 327–334, 2004.
[5] J. Srinivasan, *et.al.* The case for lifetime reliability-aware microprocessors. In *Proc. ISCA*, pp. 276–287, 2004.
[6] J. Blome, *et.al.* Self-calibrating online wearout detection. In *Proc. MICRO*, pp. 109–122, 2007.
[7] JEP122D, Failure mechanisms and models for semiconductor devices. *JEDEC Standard*, 2008.
[8] Y. Lee, *et.al.* Prediction of logic product failure due to thin-gate oxide breakdown. In *Proc. IRPS*, pp. 18–28, 2006.
[9] K. Chopra, *et.al.* A statistical approach for full-chip gate-oxide reliability analysis. In *Proc. ICCAD*, pp. 698–705, 2008.
[10] D. Brooks, *et.al.* Wattch: a framework for architectural-level power analysis and optimizations. In *Proc. ISCA*, pp. 83–94, 2000.
[11] HotSpot, http://lava.cs.virginia.edu/HotSpot/
[12] J. Sune and E. Y.Wu. Statistics of successive breakdown events in gate oxides. *IEEE Electron Device Letter*, 24(4):272–274, 2003.
[13] J. Stathis. Physical and predictive models of ultra thin oxide reliability in cmos devices and circuits. *IEEE TDMR*, 1:43–59, 2001.
[14] E. Wu, *et.al.* Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin oxides. *Solid-State Electro.*, 46(11):1787–1798, 2002.
[15] R. Degraeve, *et.al.* Temperature acceleration of oxide breakdown and its impact on ultra-thin gate oxide reliability. In *Proc.VLSIT*, pp. 59 – 60, 1999.
[16] H. Chang, *et.al.* Statistical timing analysis considering spatial correlations using a single pert-like traversal. In *Proc. ICCAD*, pp. 621–625, 2003.
[17] Z. Chen. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Stat. & Prob. Lett.*, 49(2):155–161, 2000.
[18] B. C. Kuo. *Automatic Control Systems*. Prentice Hall, 1995.