Estimating Circuit Aging due to BTI and HCI using Ring-Oscillator-Based Sensors

Deepashree Sengupta and Sachin S. Sapatnekar

***Abstract*—The performance of nanometer-scale circuits is adversely affected by aging induced by Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI). Both BTI and HCI impact transistor electrical parameters at a level that depends on the operating environment and usage of the circuit. This paper presents a novel method, using on-chip sensors based on ring oscillators (ROSCs), to detect the delay shifts in circuits as a result of aging. Our method uses presilicon analysis of the circuit to compute calibration factors that can translate BTI- and HCI-induced delay shifts in the ROSC to those in the circuit of interest. Our simulations show that the delay estimates are within 1% of the true values from presilicon analysis. Further, for post-silicon analysis, a refinement strategy is proposed where sensor measurements can be amalgamated with infrequent online delay measurements on the monitored circuit to partially capture its true workloads. This leads to about 8% lower delay guardbanding overheads compared to the conventional methods as demonstrated using benchmark circuits.**

***Index Terms*—Aging, Bias Temperature Instability, Hot Carrier Injection, Ring Oscillators, Static Timing Analysis.**

I. Introduction

With continued scaling, the susceptibility of nanometer-scale transistors to aging-related wear-out phenomena has increased significantly [1]. These aging effects cause transistor parameters to shift from their nominal values over time, resulting in a gradual degradation of circuit performance. If the extent of circuit degradation can be correctly sensed, appropriate compensation techniques can be applied to ensure reliable circuit operation. In this work, we propose a novel method to estimate aging in circuits using surrogate sensors.

We consider the two major degradation mechanisms that cause parametric delay shifts in transistors: bias temperature instability (BTI) [2], [3] and hot carrier injection (HCI) [4]. These mechanisms affect the transistor drive current by degrading the threshold voltage (V_{th}) and mobility (μ) under the voltage and temperature stress experienced during circuit operation. While BTI is partially reversible on the removal of stress, HCI is an irreversible effect. The long-term degradation due to BTI depends on the average duty cycle of the stressing signal, or the signal probability (SP), which is the probability that the signal is at the logic high level. Degradation due to HCI, on the other hand, occurs only when the device switches and hence HCI degradation depends on the time spent in switching, which is proportional to the signal activity factor (AF), i.e., the ratio of average number of signal transitions to clock transitions, and the clock frequency. In practice, most devices in a circuit tend to switch infrequently, and therefore HCI is often dominated by BTI [5]. However, over long periods, HCI grows at a faster rate than BTI, and therefore its effects can be noticeable for long-lifetime parts [6].

At a fixed supply voltage, V_{dd}, since the degradation in each transistor adversely affects its delay, the overall effect of aging is to reduce the maximum operating frequency, $\mathcal{F}_{Max}$, of a circuit with time. At the presilicon design phase, the foreknowledge of the average workload of a circuit in the field is often unavailable. Hence the schemes deployed at this phase, provide protective, albeit pessimistic, guardbands over $\mathcal{F}_{Max}$ of the circuit so that it is guaranteed to work under all operating conditions throughout its lifetime. Since both BTI and HCI aging depend on the average signal probability and activity factor (SPAF) of the stressing signals, it is common practice to choose pessimistic SPAF values for every transistor within the circuit to mimic the worst-case workload [7].

At the post-silicon stage, to ensure that a chip meets its timing requirements over its lifetime, various compensation techniques are employed during its field operation, such as clock frequency adjustment, V_{dd} scaling, and body bias modification [8]–[10]. These techniques typically use data from surrogate sensors, built in at the presilicon phase and tested at the post-silicon stage, to adaptively provide on-the-fly compensation to mitigate the effects of aging. These sensors range from simple inverter-chain-based circuits [11]–[15] to circuits based on representative critical paths [16], [17].

To a limited extent, surrogate sensors may successfully capture the environment faced by the circuit, e.g., if they are placed close to the circuit and have a similar connection to the power grid, they can capture the thermal and supply voltage environment, and undergo similar shifts due to systematic or spatially-correlated process variations. However, since these sensors are mere surrogates, they are unable to reflect aging in the circuit with complete accuracy. This inability arises due to the structural differences between the near-critical paths of the circuit under test (CUT) and the sensor. At the device level, the transistors in the sensor experience different stressing input patterns as compared to the CUT, and also have different delay sensitivities to aging shifts, due to which they age differently. At the circuit level, the number of near-critical paths in the CUT, which could potentially become critical under aging, is typically much larger than those in the sensor. A ring oscillator sensor has just one path, and although it is possible to build surrogate sensors based on representative critical path circuits (RCPs) [16], [17], their design overhead is significant. Additionally, the cost of constructing RCP circuits that could cover a sufficient number of critical paths could be onerous.

In this paper, we aim to infer delay shifts due to BTI- and HCI-induced aging in a CUT based on delay shifts

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

measured from ring-oscillator-based (ROSC-based) sensors. ROSC-based sensors are widely used because they are cheap, compact, and can be easily replicated many times within a chip. Specifically, we use the sensors in [12], which can separately measure the contribution of BTI and HCI to the delay shift of the ROSC sensor.

We propose two post-silicon schemes to estimate the delay degradation of a CUT. The first scheme translates measurement from surrogate aging sensors to circuit delay degradations using a look-up table (LUT). The second amalgamates these sensor measurements with very infrequent measurements performed directly on the CUT, and uses the results of these measurements to update the LUT. These updated LUT values are then used in conjunction with inexpensive sensor measurements to infer the delay of the CUT. While the first scheme is very easy to implement with a small increase in design effort, the second one is more accurate, albeit at the cost of increased complexity due to CUT delay measurement and LUT update circuitry. Preliminary versions of this work appeared in [18] and [19], where we considered only BTI-induced aging.

For the first scheme, we begin with a new *Upper bound on $\mathcal{F}_{Max}$* (UofM) model that estimates a safe $\mathcal{F}_{Max}$ for an aging CUT. This model accounts for the possibility that critical paths may change over the lifetime of a chip due to nonuniform delay degradation on various circuit paths, and finds an envelope for the CUT delay that provides a tight upper bound on the circuit delay. Next, we leverage the UofM model to present a novel approach for inferring the delay degradation of the CUT based on data from the on-chip aging sensors. Our scheme involves an initial presilicon characterization that uses a compact on-chip LUT to determine calibration factors. We call these the *degradation ratios*, ξ_B^{CUT} and ξ_H^{CUT}, which translate the sensor measurement data to CUT delay degradation under BTI and HCI, respectively.

While the UofM model is appropriate for presilicon aging estimation, its assumptions of worst-case aging can be quite pessimistic. We quantify the level of pessimism on a set of representative benchmark circuits and then present our second post-silicon scheme that performs infrequent measurements on the CUT and updates the calibration factors in the LUT for significantly tighter aging estimates.

The primary problem with measuring a circuit directly to estimate aging is that its normal operation must be interrupted in order to enable these measurements, which requires scheduled downtime and results in frequent but undesirable, system-level disruptions. The advantage of our methodology is that our first approach requires no such downtime, while our second approach requires minimal disruption to normal circuit operation, requiring the circuit to be measured 2-4 times over 10 years. Additionally, testing the simple ROSC circuit is fast, which causes minimal aging effects on the ROSC sensors [12].

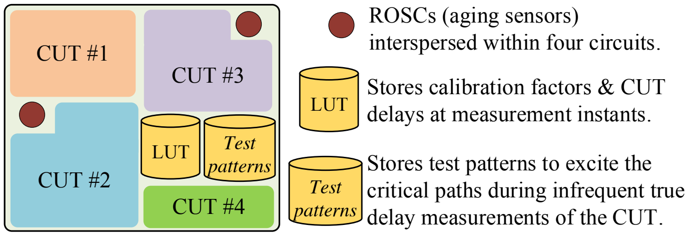

CUT #1
CUT #3
CUT #2
LUT
Test patterns
CUT #4
ROSCs (aging sensors) interspersed within four circuits.
LUT
Stores calibration factors & CUT delays at measurement instants.
Test patterns
Stores test patterns to excite the critical paths during infrequent true delay measurements of the CUT.

Fig. 1. ROSCs interspersed within multiple circuits along with LUT of degradation ratio and test pattern storage block for direct CUT measurement.

The circuit modifications associated with the proposed schemes are depicted in Fig. 1, which shows multiple ROSC sensors interspersed within four circuits. Our scheme uses the silicon odometer [12] ROSC sensor, which uses the notion of beat frequencies to measure delay variations in the ROSC to a very high degree of precision. This scheme permits the change in the period due to BTI and HCI to be measured easily and on-the-fly. The number and location of the ROSCs within a chip is a user input, and the granularity at which they are deployed reflects a trade-off between area overhead and accuracy. The degradation ratios of the CUT, and their true delays at certain instants during the CUT lifetime, are stored in the on-chip LUT. The true delay measurement circuitry is symbolically depicted by the *Test patterns* block; note that this block is only used by the second scheme, which uses these patterns to apply infrequent tests to the CUT, and is removed for the first UofM-based scheme. Under the applied test patterns, one of several existing schemes can be used here to measure the runtime delay of a circuit such as the Path-RO [20], delay shift circuits [21] [22], or by the techniques described in [9]. The degradation ratios can be recalibrated offline on the processor itself (in software) when the circuit is tested for its true delay, assuming that the processor has an arithmetic and logic unit.

The overall flow of the proposed two-step framework can be summarized as follows:

(1) Our first scheme obtains the degradation ratios from presilicon characterization, and uses them to translate the ROSC measurements to CUT delay degradation under aging.

- We obtain an envelope of the CUT delay under aging by performing aging-aware static timing analysis (STA) assuming worst-case workload on the CUT.
- Using the above envelope, and aging-aware STA on the ROSC sensors, we obtain the degradation ratios, and store them in the on-chip LUT.
- After the chip has been deployed in field operation, we probe the ROSC sensor from time to time to observe its delay degradation separately due to BTI and HCI.
- Multiplying the ROSC delay degradation with the corresponding degradation ratios from the LUT produces the CUT delay degradation, which is an indicator of its aging due to BTI and HCI.

(2) Our second scheme is built upon the first one, where the ROSC measurements are still used to infer CUT degradation, however, the degradation ratios are now updated in the LUT (although infrequently) after obtaining the CUT delay at prespecified time instants, called the *measurement instants*.

- The degradation ratios stored in the LUT from presilicon analysis and postsilicon ROSC delay degradation data are used to obtain CUT delay till the first measurement instant, using our first scheme.
- At each measurement instant, the true delay of the CUT is measured using a separate dedicated circuit, and stored

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

TABLE I
LIST OF THE FREQUENTLY USED SYMBOLS IN THE PAPER.

Symbol	Explanation
$\Delta V_{th_n}(t)$ ($\Delta V_{th_p}(t)$)	Threshold voltage degradation of an NMOS (PMOS) from time, t_0 to t, where t_0 is the time when observation begins.
$f(t)$ ($g(t)$)	Temporal dependence of threshold voltage degradation by Reaction Diffusion (Charge Trapping) model.
$\Delta f(t)$ ($\Delta g(t)$)	$f(t) - f(t_0)$ ($g(t) - g(t_0)$)
$D^p_{pre,B}(t)$ ($D^p_{pre,H}(t)$)	Presilicon estimate of delay of path, p, of a circuit under BTI (HCI) aging alone.
$D^C(t)$ ($D^C_{\text{UofM}}(t)$)	Actual value (UofM estimate) of delay of the circuit, C, at time, t, under aging.
K_B (K_H)	Proportionality constants for the dependence of delay change to $\Delta f(t)$ ($\Delta g(t)$), called the K_B (K_H) value of the aging curve.
K^X_B (K^X_H)	K_B (K_H) value of the aging curve of the structure defined by X which could be a cell, a path, or a circuit.
ξ^C_B (ξ^C_H)	Degradation ratio of the circuit, C, corresponding to BTI (HCI) aging.
$\Delta D^C_{post}(t)$	Estimated delay degradation of the circuit, C, from time, t_0, using postsilicon ROSC measurements and degradation ratios.
t_{m_j}, $j = 0, \cdots, N-1$	Set of N measurement instants at which the circuit delay is measured and the degradation ratios are recalibrated.
$D^C_{re}(t)$	Recalibrated upper bound on delay of the circuit, C, taking true delay measurements into account.
$K^X_{j,B}$ ($K^X_{j,H}$)	Recalibrated K_B (K_H) value of the aging curve of X (cell, path, or circuit) after measurement instant, t_{m_j}.
$\xi^C_{j,B}$ ($\xi^C_{j,H}$)	Recalibrated degradation ratio of the circuit, C, corresponding to BTI (HCI) aging, after measurement instant, t_{m_j}.
$\Delta D^C_{post,j}(t)$	Estimated delay degradation of the circuit, C, from time, t_{m_j}, using ROSC measurements and the recalibrated degradation ratios.

in the LUT, along with updating the degradation ratios.

- Multiplying the ROSC delay degradation with the most recent degradation ratios from the LUT produces the CUT delay degradation, and adding that to the true delay of the CUT at the previous measurement instant produces a more accurate estimate of aging in the CUT.

A detailed overview of the first scheme is presented in Sec. III, while the second one is explained in Sec. IV. We outline a brief background on BTI- and HCI-induced aging, and the resulting delay degradation in circuits in Sec. II. Sec. V demonstrates the experimental setup and results, and we conclude in Sec. VI.

The frequently used symbols in the paper are summarized in a list in Table I.

II. BACKGROUND ON BTI AND HCI

Under BTI, a PMOS (NMOS) device is stressed when its gate voltage is negative (positive), leading to negative (positive) BTI, or NBTI (PBTI), while under HCI, a transistor is stressed only while it switches. In this section, we describe models for BTI and HCI aging, their impact on delay degradation of individual transistors, and on larger circuits.

In the discussion below, we will assume that the observation time for the circuit starts at time, t_0, and continues until the lifetime of the circuit, t_f.

A. BTI-induced aging

The precise mechanism of BTI is a matter of debate within the research community. Two candidates have emerged: the reaction-diffusion (RD) model [2] and the charge trapping (CT) model [23]. The threshold voltage shift is a cumulative effect of multiple cycles of stress and recovery. BTI is independent of the frequency of the stressing signal for frequencies higher than 1Hz, and only depends on its average duty cycle [3]. In general, at time, t, the threshold voltage shift, $\Delta V_{th_x}(t)$, $x \in \{n, p\}$, due to BTI in an NMOS or PMOS device can be modeled as:

$$\Delta V_{th_x}(t) = C_x(f(t_{\text{st}}) - f(t_{\text{st},0})) = \psi_{B_x}(f(t) - f(t_0)) \quad (1)$$

where C_x is a constant dependent on the process, voltage and temperature (PVT) conditions of the device, $f(.)$ is a function that represents the temporal dependence of BTI aging, and the terms, $t_{\text{st},0}$ and t_{st}, refer to the effective stressing time after an elapsed time of t_0 and t, respectively. The effective stressing times are given by $t_{\text{st}} = \alpha t$ (and $t_{\text{st},0} = \alpha t_0$), where α is the stress probability for the device. Since a PMOS device is stressed when its input is at logic low level, $\alpha = 1 - s$, where s is the SP of the input signal. Similarly, for an NMOS device, $\alpha = s$, since it is stressed when its input is at logic high level. The $f(t_{\text{st},0})$ term thus relates to the aging of the chip that is built in at time, t_0. We can absorb the effect of the signal probability into ψ_{B_x} along with other PVT dependent parameters in C_x, and we note that $f(t)$ is purely dependent on the age of the circuit, t. In principle, $f(.)$ could be different for PMOS and NMOS devices, but these are experimentally observed to be similar, as documented in design manuals and the published literature [24]. Typically, $f(t)$ can assume either of the following forms based on the two models of BTI:

$$f(t) = \begin{cases} t^{n_1} & \text{, under the RD model} \\ a + b \log t & \text{, under the CT model} \end{cases} \quad (2)$$

where $n_1 \sim 0.16$ [25], and a and b are positive constants defined in [23]. Our analysis in this paper is designed to be general enough to be applicable on either form of $f(t)$. Although the V_{th}-shifts through multiple stress-recovery cycles are not monotonic, Eq. (1) captures the envelope of the delay function, including BTI recovery effects under AC stress.

B. HCI-induced aging

At contemporary technology nodes, HCI affects NMOS devices more severely than PMOS [4], [26]. HCI occurs when carriers in the channel, subjected to a lateral electric field, gain sufficient energy and momentum to break the barriers of surrounding dielectric, such as the gate and sidewall oxides. A recent energy-driven framework for HCI stress proposes that carriers with sufficient energy can result in interface-state generation by impact ionization at the Si-SiO$_2$ interface

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

directly without being injected into the gate oxide [4]. This leads to a gradual degradation in various electrical parameters of the transistors, thus affecting the circuit performance.

Based on [13], [27], we model HCI aging by expressing the drive current reduction as equivalent threshold voltage degradation, $\Delta V_{th_n}(t)$, of an NMOS after time, t, as:

$$\Delta V_{th_n}(t) = C_H \exp\left(\frac{E_{ox}}{E_0} - \frac{\phi_{it}}{q\lambda E_m}\right)(g(t_{\text{st}}) - g(t_{\text{st},0})) \tag{3}$$

where C_H and E_0 are process dependent parameters, E_{ox} is the vertical field, ϕ_{it} is the trap generation energy, q is the electronic charge, λ is the hot electron mean free path, t_{st} and $t_{\text{st},0}$ are the effective stressing times, $g(.)$ is a function that encapsulates the temporal dependence of HCI aging, and E_m is the lateral electric field, given by:

$$E_m = \frac{V_{ds} - V_{dsat}}{L_{\mathit{eff}}} \tag{4}$$

$$\text{where } V_{dsat} = \frac{(V_{gs} - V_{th} + \frac{2k_BT}{q})L_{\mathit{eff}}E_{sat}}{V_{gs} - V_{th} + \frac{2k_BT}{q} + A_{bulk}L_{\mathit{eff}}E_{sat}} \tag{5}$$

The parameter, L_{eff}, is the effective channel length, T is the temperature, k_B is Boltzmann's constant, and A_{bulk} and E_{sat} are process-dependent constants defined in [28].

The effective stressing time, t_{st}, corresponding to HCI aging depends on the number of switching events experienced by the transistor, given by ($\text{AF} \cdot \mathcal{F}_{clk} \cdot t$), where AF is the activity factor for the transistor, $\mathcal{F}_{clk}$ is the clock frequency, and t is the elapsed time. During each of these switching events, the transistor is stressed during the time that the input signal makes its transition, given by the slew, t_{slew}. Hence, we can write $t_{st} = (\text{AF} \cdot \mathcal{F}_{clk} \cdot t)t_{slew}$, and this relation converts $g(t_{\text{st}})$ to a function $g(t)$ of the elapsed time. Similarly, $g(t_{\text{st},0})$ can be converted to $g(t_0)$ as well. We thus absorb the effect of the switching activity and the time-independent parameters into ψ_{H_n} to rewrite $\Delta V_{th_n}(t)$ as:

$$\Delta V_{th_n}(t) = \psi_{H_n}(g(t) - g(t_0)) \tag{6}$$

From experimental models [4], $g(t)$ is typically of the form:

$$g(t) = t^{n_2} \tag{7}$$

where $n_2 \sim 0.5$.

C. Combined effects of BTI and HCI on delay degradation

We denote the change in the underlying trend functions for BTI and HCI as:

$$\Delta f(t) = f(t) - f(t_0) \tag{8}$$

$$\Delta g(t) = g(t) - g(t_0) \tag{9}$$

Between time, t_0, to t, we can represent the shift in the delay, $D(t)$, of a logic gate as:

$$\Delta D(t) = \sum_{i \in \text{NMOS}} S_{n,i}\Delta V_{th_{n,i}}(t) + \sum_{i \in \text{PMOS}} S_{p,i}\Delta V_{th_{p,i}}(t) \tag{10}$$

where the two summations are taken over all NMOS and PMOS transistors in a gate. Here, the prefix, Δ, denotes a change in the quantity that succeeds it, $V_{th_{n,i}}(t)$ ($V_{th_{p,i}}(t)$) is the threshold voltage of the i^{th} NMOS (PMOS) transistor in the gate, and $S_{x,i} = \left.\frac{\partial D}{\partial V_{th_{x,i}}}\right|_{t_0}$ for $x \in \{n, p\}$ is the sensitivity of transistor, i, to threshold voltage shifts. From Eqs. (1) and (6), we have:

$$\begin{aligned} \Delta V_{th_{n,i}}(t) &= \psi_{B_{n,i}}\Delta f(t) + \psi_{H_{n,i}}\Delta g(t) \\ \Delta V_{th_{p,i}}(t) &= \psi_{B_{p,i}}\Delta f(t) \end{aligned} \tag{11}$$

Therefore, we can rewrite Eq. (10) as:

$$\Delta D(t) = K_B\, \Delta f(t) + K_H\, \Delta g(t) \tag{12}$$

where $K_B = \left(\sum_{i \in \text{NMOS}} S_{n,i}\psi_{B_{n,i}} + \sum_{i \in \text{PMOS}} S_{p,i}\psi_{B_{p,i}}\right)$ and $K_H = \sum_{i \in \text{NMOS}} S_{n,i}\psi_{H_{n,i}}$. Thus, under fixed stress conditions of temperature, V_{dd}, SP, and AF, the delay is a function of time, and is easily computed if all the sensitivity values, $S_{n,i}$ and $S_{p,i}$, have been characterized for each gate.

III. Delay estimation and aging prediction

Our goal is to estimate a safe value of the maximum frequency of operation, $\mathcal{F}_{Max}$, of the CUT under aging, using data from on-chip ROSCs. We obtain this estimate by determining degradation ratios, ξ_B^{CUT} and ξ_H^{CUT}, that multiply the delay degradation of nearby ROSC test structures to estimate the delay shift in the CUT. Our initial analysis, described in this section, performs presilicon analysis to determine the values for these degradation ratios, which are held constant over the lifetime of the circuit and stored in the LUT in Fig. 1. This scheme does not use the *Test patterns* block in the figure.

We begin by observing that the rate at which a path ages depends on how it is stressed and on its sensitivity to stress. Due to this, the critical path of a CUT may change over its lifetime owing to the nonuniform delay degradation on its near-critical paths. Fig. 2 depicts possible aging trajectories for several near-critical paths of a CUT, C. The delay, $D^C(t)$, of the CUT is the maximum among the delays of the near-critical paths, and this is seen to be a piecewise-smooth curve. In contrast, the ROSC has a single path that ages along a constant profile through its lifetime and has a smooth trajectory, $D^R(t)$, similar to any of the path delays in Fig. 2.

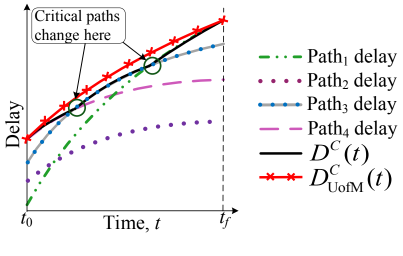


Fig. 2. CUT delay as maximum of path delays under aging.

The above example indicates a primary difficulty in using a ROSC-based sensor to predict the delay degradation of the CUT, since it is nontrivial to develop a simple one-to-one functional relationship between a smooth trajectory (corresponding to the ROSC delay, $D^R(t)$) and a nondifferentiable function (which characterizes the delay, $D^C(t)$, of the CUT).

To overcome this, we first obtain a pessimistic and continuously differentiable presilicon bound, $D^C_{\text{UofM}}(t)$, of the

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

CUT delay, as illustrated in Fig. 2. We refer to this as the *Upper bound on $\mathcal{F}_{Max}$* (UofM) model. To ensure pessimism, $D^C_{\text{UofM}}(t)$, must lie above $D^C(t)$, $\forall t \in [t_0, t_f]$, so that if $D^C_{\text{UofM}}(t)$ meets the timing requirements throughout the lifetime, then so does $D^C(t)$. Next, we find a relation between the ROSC delay and the UofM delay to obtain the degradation ratios to estimate CUT delay degradation from the ROSC data.

In Sec. III-A, we discuss the presilicon characterization of the CUT and the ROSC to compute the degradation ratios. Next, in Sec. III-B, we outline the methodology of post-silicon aging estimation in the CUT from the ROSC measurements using these ratios, and in Sec. III-C, we examine the validity of using degradation ratios characterized at the presilicon stage, since post-silicon operating conditions can be different due to PVT variations, dynamic voltage scaling, and supply gating.

A. Presilicon circuit characterization

UofM model: We first present a theorem to obtain an expression for a differentiable function that is an upper bound for the maximum of n functions, each of the form similar to Eq. (12).

Theorem 1 *In the interval, $[t_0, t_f]$, consider a set of monotonically increasing functions, $x_1(t)$, $\cdots$, $x_n(t)$, such that $x_i(t) = x_i(t_0) + \theta^i_1 \Delta f(t) + \theta^i_2 \Delta g(t)$, with θ^i_1, $\theta^i_2 \geq 0$, where $\Delta f(t) = f(t) - f(t_0)$ and $\Delta g(t) = g(t) - g(t_0)$. Then for $f(t) = t^{n_1}$ or $a + b \log t$, and $g(t) = t^{n_2}$ with $a, b > 0$, and $1 > n_2 > n_1 > 0$, an upper bound on the maximum of these functions is given by another function, $y(t)$, of similar form:*

$$y(t) = x_M(t_0) + \theta^M_1 \Delta f(t) + \theta^M_2 \Delta g(t) \tag{13}$$

where $x_M(t)$ is the maximum envelope of the $x_i(t)$ functions, such that $x_M(t_0) = \max_{i \in \{1,\cdots,n\}}(x_i(t_0))$.

The coefficients, θ^M_1 and θ^M_2, are obtained from the θ^i_1 values, and by evaluating $x_M(t)$ at time instants, $t = t_0$ and t_f, and are defined as:

$$\theta^M_1 = \max_i(\theta^i_1) \tag{14}$$

$$\theta^M_2 = \frac{\Delta x_M(t_f) - \theta^M_1 \Delta f(t_f)}{\Delta g(t_f)} \tag{15}$$

where $\Delta x_M(t) = x_M(t) - x_M(t_0)$.

A brief outline of the proof is that the UofM model is constructed as a curve of the form of Eq. (12) that matches the piecewise-smooth maximum function, $x_M(t)$, at $t = t_0$ and $t = t_f$, and lies above it at all other points, $t \in (t_0, t_f)$. The detailed proof is deferred to Appendix A.

To map the results of this theorem to our problem, we represent the delay of each near-critical path, p_i, of the CUT in the form of $x_i(t)$, and use Theorem 1 to determine the UofM delay bound. Note that the restrictions on a, b, n_1, and n_2 are easily satisfied by typical BTI/HCI models. The evaluation of the upper bound in Theorem 1 requires the values of $x_M(t_0)$, $x_M(t_f)$, and θ^i_1 to be determined[1]. For the problem at hand, evaluating $x_M(t)$ at time instants, $t = t_0$ and $t = t_f$, in the theorem is equivalent to obtaining the presilicon circuit delay, $D^C_{pre}(t)$, at these two instants by performing STA on C. Obtaining θ^i_1 is equivalent to computing the K_B value of the delay trajectory of p_i. To obtain K_B (K_H) value of p_i, we simply evaluate the delay of p_i at $t = t_0$ and t_f under BTI (HCI) aging alone for specific workload conditions, and compute $K^{p_i}_B$ and $K^{p_i}_H$ as:

$$K^{p_i}_B = \frac{D^{p_i}_{pre,B}(t_f) - D^{p_i}_{pre}(t_0)}{\Delta f(t_f)} \tag{16}$$

$$K^{p_i}_H = \frac{D^{p_i}_{pre,H}(t_f) - D^{p_i}_{pre}(t_0)}{\Delta g(t_f)} \tag{17}$$

where $D^{p_i}_{pre,B}(t)$ $\left(D^{p_i}_{pre,H}(t)\right)$ is the evaluated delay of p_i at t due to BTI (HCI) aging alone, and $D^{p_i}_{pre}(t_0)$ is the presilicon delay at time t_0.

To summarize, we obtain the smooth upper bound, $D^C_{\text{UofM}}(t)$, on $D^C(t)$ by using any timing analysis tool (home-grown or commercial) to perform two STA evaluations, at times, t_0 and t_f, on the CUT using the aging models for BTI and HCI from Sec. II. Since we must perform these presilicon STA runs to account for all parts in the field excited with any SP or AF value, the specific workload conditions correspond to choosing these values pessimistically. For BTI analysis, we assume the worst-case SP of 1 for each gate input; similarly, for HCI analysis, we assume an AF of 1. For each near-critical path, p_i, a pair of timing evaluations under BTI aging alone can be used to compute the $K^{p_i}_B$ value for the path using Eq. (16). Finally, we use Theorem 1 to characterize the constants, K^C_B and K^C_H, in $D^C_{\text{UofM}}(t)$, similar to θ^M_1 and θ^M_2 in $y(t)$ in the theorem, to obtain:

$$\Delta D^C_{\text{UofM}}(t) = D^C_{\text{UofM}}(t) - D^C_{\text{UofM}}(t_0) = K^C_B\ \Delta f(t) + K^C_H\ \Delta g(t) \tag{18}$$

where $K^C_B \Delta f(t)$ and $K^C_H \Delta g(t)$ are the aging contributions due to BTI and HCI, respectively, to the total delay degradation from $t = t_0$.

Degradation ratios from ROSC and CUT delay analysis: A ROSC is a chain of an odd number, $2l + 1$, of inverters connected in a closed loop. Assuming, for simplicity, that each inverter has a rise delay of $d_r(t)$ and a fall delay of $d_f(t)$, the period of the ROSC is well known to be $(2l+1)(d_r(t)+d_f(t))$. We refer to the period of a ROSC, R, as its delay, $D^R(t)$.

Since the ROSC has 50% signal probability and toggles on every clock transition, the SP values at each of its gate inputs is 0.5 and its AF is 1. At the presilicon stage, the aging trends for the ROSC can be characterized to separately evaluate K^R_B and K^R_H. This requires an aging-aware timing analysis of the single critical path of the ROSC, first assuming only BTI, and then only HCI aging, each at $t = t_0$ and $t = t_f$, similar to Eqs. (16) and (17). The presilicon estimate of the delay degradation, $\Delta D^R_{pre}(t)$, of the ROSC is obtained in the form of Eq. (12), as:

$$\Delta D^R_{pre}(t) = D^R_{pre}(t) - D^R_{pre}(t_0) = K^R_B \Delta f(t) + K^R_H \Delta g(t) \tag{19}$$

Based on the K_B and K_H values computed to characterize Eqs. (18) and (19), we compute the degradation ratios, ξ^C_B

[1] Superficially, it may appear that Eqs. (14) and (15) are independent of θ^i_2, but the influence of this parameter is hidden from view in $\Delta x_M(t_f)$.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

and ξ_H^C, of the CUT, C, corresponding to BTI and HCI aging, respectively, as:

$$\xi_B^C = \frac{K_B^C}{K_B^R} \qquad \xi_H^C = \frac{K_H^C}{K_H^R} \tag{20}$$

Note that the degradation ratios above depend purely on the ratios of the K_B and K_H values and are independent of time.

B. Post-silicon aging estimation in the CUT from ROSC data

The values of the degradation ratios for each CUT, with respect to its associated ROSC, are stored in the on-chip LUT depicted in Fig. 1. To use this LUT in the post-silicon context at time t, we will separately measure the values of delay shifts, $\Delta D_B^R(t)$ and $\Delta D_H^R(t)$, due to BTI and HCI, respectively, of the silicon odometer ROSC [12]. We infer $K_{post,B}^R = \Delta D_B^R(t)/\Delta f(t)$ and $K_{post,H}^R = \Delta D_H^R(t)/\Delta g(t)$ from the silicon odometer ROSC; recall that this sensor can separately measure its own BTI- and HCI-induced degradation. Based on this measurement, we estimate the post-silicon CUT delay degradation, $\Delta D_{post}^C(t)$, in a manner similar to Eq. (18) as:

$$\begin{aligned}\Delta D_{post}^C(t) &= K_{post,B}^C\ \Delta f(t) + K_{post,H}^C\ \Delta g(t)\\ &= K_{post,B}^R\ \xi_B^C \Delta f(t) + K_{post,H}^R\ \xi_H^C \Delta g(t)\end{aligned} \tag{21}$$

where $K_{post,B}^C$ and $K_{post,H}^C$ correspond to the unknown K_B and K_H values, respectively, of the CUT delay trajectory at the post-silicon stage, which we want to infer from the ROSC measurements. The above equation seems to imply that ξ_B^C and ξ_H^C, defined in Eq. (20), can also be formulated as:

$$\xi_B^C = \frac{K_{post,B}^C}{K_{post,B}^R} \qquad \xi_H^C = \frac{K_{post,H}^C}{K_{post,H}^R}$$

In other words, the degradation ratios from the presilicon stage are the same at the post-silicon stage, in spite of both being characterized by different operating conditions. This assumption is correct for all practical purposes, as will be explained in the next section.

C. The effects of operating conditions on degradation ratios

During circuit operation in the field, when the above ROSC measurements are taken and the CUT delay shift is estimated, both the CUT and the ROSC age under conditions that are different from those during presilicon estimation. We now critically examine the validity of using presilicon degradation ratios under these conditions, considering the effect of each factor that differs between the presilicon and post-silicon phases: specifically, the SP and AF for the circuit; systematic process variations; temperature, T, and supply voltage, V_{dd}, including supply gating and dynamic voltage scaling.

Circuit SP and AF: The precharacterized relation in Eq. (20) uses a worst-case SP and AF scenario for aging, and therefore provides a pessimistic estimate of the CUT delay.

Process variations: The proximity of the ROSC and the CUT ensures that both have similar systematic variations in the process parameters (length, width, oxide thickness, and other critical dimensions). The dominant systematic variations [29] [30] are thus, very similar for both the CUT and the ROSC. As a result, the process dependence of K_B and K_H values in Eq. (12) are also similar for both the CUT and the ROSC, and hence the degradation ratios, ξ_B^C and ξ_H^C, which are the ratios of K_B and K_H of the CUT to that of the ROSC, are practically independent of the process variations. The effect of the random variations is also minimized for a ROSC and CUT with multiple stages [11] [31], considered in this work.

Voltage and temperature variations: To investigate the impact of V_{dd} and T on these ratios, it is necessary to analyze K_B and K_H, as given in Sec. II-C. We can rewrite the (V_{dd}, T)-dependent parts in K_B and K_H from [32] and Eq. (3) in Sec. II-B, respectively, as:

$$K_B = \sum_{x\in\{n,p\}} S_x \psi_{B_x} = \Gamma_B P_B \exp\left(\frac{2E_{ox}}{E_0} - \frac{E_a}{k_B T}\right)\mathcal{S}(V_{dd},T) \tag{22}$$

$$K_H = \sum_{x\in\{n\}} S_x \psi_H = \Gamma_H P_H \exp\left(\frac{E_{ox}}{E_0} - \frac{\phi_{it}}{q\lambda E_m}\right)\mathcal{S}(V_{dd},T) \tag{23}$$

where P_B and P_H are the combined process-dependent terms, Γ_B and Γ_H indicate the effect of average SPAF conditions on BTI and HCI aging, and $\mathcal{S}(V_{dd},T)$ is a general function representing the dependence of the delay sensitivities to the operating conditions, which can be assumed to be similar for both PMOS and NMOS. The exact form of $\mathcal{S}$ is not necessary for our analysis.

For BTI, the dependence on V_{dd} is embedded within the terms, $\mathcal{S}(V_{dd},T)$ and $E_{ox} = \frac{V_{dd}-V_{th}}{T_{ox}}$, while the relationship with T is explicitly visible in the term, $\exp\left(-\frac{E_a}{k_B T}\right)$, in Eq. (22). For HCI degradation, the dependence on V_{dd} is visible in $\mathcal{S}(V_{dd},T)$ as well as E_m in Eq. (4), while the trend with T is implicitly hidden in the V_{dsat} term in Eq. (5) that is referenced in the equation for E_m.

The gate delay shifts can be obtained by combining Eqs. (12), (22), and (23). For any path, X, which could be a near-critical path, p_i, of the CUT or the single path, r, of the ROSC, R, we can represent its actual delay degradation from t_0 due to BTI and HCI, by $\Delta D_B^X(t)$ and $\Delta D_H^X(t)$, respectively, as:

$$\Delta D_B^X(t) = K_B^X \Delta f(t) = \mathcal{K}_B^X F_B(V_{dd},T)\Delta f(t) \tag{24}$$

$$\Delta D_H^X(t) = K_H^X \Delta g(t) = \mathcal{K}_H^X F_H(V_{dd},T)\Delta g(t) \tag{25}$$

where K_B^X and K_H^X represent the K_B and K_H values of the path, X, respectively, and $\mathcal{K}_B^X$ and $\mathcal{K}_H^X$ are the (V_{dd}, T)-independent effects of adding the gate delays along X under BTI and HCI aging, respectively. The functions, $F_B(.)$ and $F_H(.)$, represent the (V_{dd}, T)-dependent terms for BTI and HCI, respectively, as:

$$F_B(V_{dd},T) = \exp\left(\frac{2E_{ox}}{E_0} - \frac{E_a}{k_B T}\right)\mathcal{S}(V_{dd},T) \tag{26}$$

$$F_H(V_{dd},T) = \exp\left(\frac{E_{ox}}{E_0} - \frac{\phi_{it}}{q\lambda E_m}\right)\mathcal{S}(V_{dd},T) \tag{27}$$

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Theorem 2 *For a given CUT, C, let p_f and p_0 be the paths that are critical at times, t_f and t_0, respectively, and r be the single path in the ROSC. Among all near-critical paths of C, let p_m be the path with the maximum value of K_B. Then,*

1) *The degradation ratio, ξ_B^C, of the CUT, C, is independent of the supply voltage, V_{dd}, and temperature, T.*
2) *The degradation ratio, ξ_H^C, has the following dependence:*

$$\xi_H^C = \frac{\Delta D^{p_f,p_0}(t_0) + \Delta\mathcal{K}_B^{p_f,p_m} F_B(V_{dd},T)\Delta f(t_f)}{\mathcal{K}_H^r F_H(V_{dd},T)\Delta g(t_f)} + \frac{\mathcal{K}_H^{p_f}}{\mathcal{K}_H^r} \tag{28}$$

where $\Delta D^{p_f,p_0}(t_0) = D^{p_f}(t_0) - D^{p_0}(t_0)$ is the difference in the delays of paths, p_f and p_0, at $t = t_0$, $\Delta\mathcal{K}_B^{p_f,p_m} = \mathcal{K}_B^{p_f} - \mathcal{K}_B^{p_m}$.

The factors, $\mathcal{K}_B^X$ and $\mathcal{K}_H^X$, for $X \in \{p_f, p_m, r\}$, are as defined in Eqs. (24) *and* (25)*, and the functions, $F_B(V_{dd},T)$ and $F_H(V_{dd},T)$, are defined in Eqs.* (26) *and* (27)*, respectively.*

The formal proof of the theorem is deferred to Appendix B for better readability.

Based on Theorem 2, ξ_B^C is independent of V_{dd} and T, while ξ_H^C is not. Hence, assuming (V_{dd}, T)-independence of ξ_H^C will incur some error in the aging estimate due to HCI, which can be bounded by a maximum value with the knowledge of the operating conditions according to Eq. (28). However, in reality, the near-critical paths have similar delay profiles (i.e., $D^{p_0}(t_0) \approx D^{p_f}(t_0)$), as well as similar aging trends (i.e., $\mathcal{K}_B^{p_f} \approx \mathcal{K}_B^{p_m}$), due to which $\xi_H^C \approx \frac{\mathcal{K}_H^{p_f}}{\mathcal{K}_H^r}$ from Eq. (28) in Theorem 2. Hence the error is negligible while considering (V_{dd}, T)-independence of ξ_H^C as well. In the rest of the paper we thus assume both ξ_B^C and ξ_H^C to be independent of V_{dd} and T.

Dynamic voltage scaling and V_{dd} gating: It is important to note that the above analysis is valid for any variation in V_{dd}, as long as the ROSC and the CUT are subjected to the same variation. Therefore, it is valid not only for fluctuations in the supply noise, which are identical due to spatial locality of a CUT and its nearby ROSC, but also for V_{dd} changes due to dynamic voltage scaling or V_{dd} gating.

Hence the degradation ratios are practically independent of PVT variations, deliberate supply voltage changes, and time, and require characterization only once for the entire lifetime of a circuit to obtain its pessimistic delay estimate under the assumption of a worst-case SPAF workload, at the post-silicon stage from ROSC measurements.

IV. Sensor recalibration and aging estimation

While the ROSC described in Sec. III can track PVT variations, voltage scaling, and V_{dd} gating in the CUT, it is inherently incapable of tracking changes in the SPAF of the CUT. Hence, the interpretation of the ROSC data must assume SPAF settings that correspond to worst-case aging in the CUT. This may result in pessimistic estimates of the circuit performance, and may underestimate circuit delays by over 10%, as documented in Sec. V.

The scheme presented in this section supplements data from the ROSC through infrequent direct measurements of the CUT, measuring its true delay in the field. The information gathered through these measurements is used to recalibrate the relationship between ROSC aging and CUT aging. In short, we combine infrequent delay measurements on the CUT with cheap and more frequent ROSC measurements, to obtain more accurate estimates of the CUT delay. This allows ROSC measurements to be personalized to individual chips in the field, accounting for the way they age, based on the specific stressing environment that the part is subjected to.

Our approach proceeds by recalibrating the degradation ratios, K_B and K_H, based on data from CUT measurements, so that ROSC measurements can be mapped more accurately on to CUT delay estimates. The modified scheme uses the LUT depicted in Fig. 1 to store the degradation ratios and the *Test Patterns* block to store the test patterns required to determine the delay degradation of the circuit.

A. Delay bounds based on post-silicon CUT measurements

We illustrate the scheme through Fig. 3. The pessimistic delay degradation trajectory over all possible workloads is given by the UofM bound, $D_{\mathrm{UofM}}^C(t)$. However, for a specific chip running a particular workload, the actual delay degradation follows the curve shown by $D_a^C(t)$; by definition, this must lie below the UofM bound. To correct this difference, the degradation ratios are modified at a set of measurement instants, corresponding to $t = t_{m_0}, \cdots, t_{m_3}$, by performing direct measurements on the circuit and appropriately recalibrating K_B^C and K_H^C to achieve a better prediction, shown by $D_{re}^C(t)$.

Up to the first measurement instant, t_{m_1}, $D_{re}^C(t)$ exactly follows the UofM bound. At this point, the bound is brought down to the measured delay, and the delay trajectory beyond this point must be predicted. Any such projection of future activity must be made without foreknowledge of the workload and therefore, the $D_{re}^C(t)$ curve must necessarily assume worst-case aging beyond t_{m_1}. The actual aging curve will lie below this bound, and at the next measurement instant, t_{m_2}, a recalibration is made to match the measured value, and so on. As a result, the $D_{re}^C(t)$ curve matches the actual aging curve, $D_a^C(t)$, more closely than the UoM curve, $D_{\mathrm{UofM}}^C(t)$.

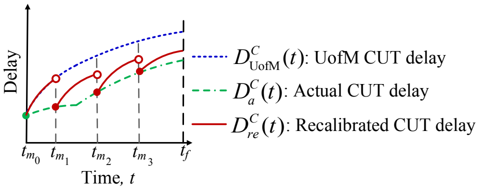


Fig. 3. Upper-bounding the CUT delay with intermediate CUT measurements.

The delay measurements may be performed on the CUT using one of several existing schemes, such as the Path-RO [20], delay shift circuits [21] [22], or by the techniques described in [9]. These techniques typically use input vectors stored in an on-chip memory, with a test controller, which fetches a pair of these vectors to perform the delay tests. The number of such patterns to excite near-critical paths of a CUT

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

to measure the worst-case delay, is sufficiently small, and the hardware overhead of the associated test controller is less than 0.01% as reported in [22]. The *Test patterns* block in Fig. 1, thus abstracts the entire circuitry that is appended to the circuit block for its true delay measurement.

The following result provides two upper bounds to obtain $D_{re}^C(t)$ for $t \in [t_{m_j}, t_{m_{j+1}}]$.

Theorem 3 *Let $\{t_{m_0}, \cdots, t_{m_{N-1}}\}$ be the N measurement instants at which the degradation ratios are recalibrated, and $t_{m_N} = t_f$, and let*

$$\Delta f_j(t) = f(t) - f(t_{m_j})$$
$$\Delta g_j(t) = g(t) - g(t_{m_j})$$

After each measurement instant, the recalibrated upper bound on the delay, or, $D_{re}^C(t)$, is obtained as follows.
For $0 \leq t < t_{m_1}$,

$$D_{re}^C(t) = D_{\text{UofM}}^C(t) \tag{29}$$

For $t_{m_j} \leq t \leq t_{m_{j+1}}, j > 1$, two upper bounds on $D_a^C(t)$ are:

$$\textbf{(I)}\ D_{re}^{C,I}(t) = D_a^C(t_{m_j}) + K_B^I \Delta f_j(t) + K_H^{II} \Delta g_j(t), \tag{30}$$

If $p_x = \{p_i \in S_{NC} | D_{pre}^{p_i}(t_{m_{j+1}}) - D_{pre}^{p_i}(t_{m_j})$ is maximized$\}$, $K_B^I = K_B^{p_x}$, $K_H^I = K_H^{p_x}$, where S_{NC} is the set of near-critical paths of C, and $D_{pre}^{p_i}(t) = D_{pre}^{p_i}(t_0) + K_B^{p_i} \Delta f(t) + K_H^{p_i} \Delta g(t)$ is the worst-case presilicon delay estimate of path $p_i \in S_{NC}$.

$$\textbf{(II)}\ D_{re}^{C,II}(t) = D_a^C(t_{m_j}) + K_B^{II} \Delta f_j(t) + K_H^{II} \Delta g_j(t) \tag{31}$$
$$K_B^{II} = \max_{p_i \in S_{NC}} (K_B^{p_i}),$$
$$K_H^{II} = \frac{[D_{\text{UofM}}^C(t_{m_{j+1}}) - D_a^C(t_{m_j})] - K_B^{II} \Delta f_j(t_{m_{j+1}})}{\Delta g_j(t_{m_{j+1}})},$$

where $K_B^{p_i}$ is the K_B value of the path p_i.

To use Theorem 3 for $j > 1$, we choose the bound that is tighter at time $t_{m_{j+1}}$. We select $K_{j,B}^C$ and $K_{j,H}^C$ as follows:

- If $D_{re}^{C,I}(t_{m_{j+1}}) \geq D_{re}^{C,II}(t_{m_{j+1}})$, then
$$K_{j,B}^C = K_B^I \qquad K_{j,H}^C = K_H^I \tag{32}$$
- Otherwise
$$K_{j,B}^C = K_B^{II} \qquad K_{j,H}^C = K_H^{II} \tag{33}$$

The proof of the upper bounds in Theorem 3 is presented in Appendix C. Hence, for $t \in [t_{m_j}, t_{m_{j+1}}]$, $j \geq 1$,

$$D_{re}^C(t) = D_a^C(t_{m_j}) + K_{j,B}^C \Delta f_j(t) + K_{j,H}^C \Delta g_j(t) \tag{34}$$

Intuitively, Case I provides one candidate for an upper bound, but if that bound exceeds the UofM prediction, Case II provides a tighter estimate.

An example of delay estimation using the current scheme for a single measurement instant (excluding t_{m_0}) is shown in Fig. 4. The two circuits, mem_ctrl and i2c, are from the IWLS'05 benchmark suite [33], and are stressed at 125°C and 1.2V. The real delay curve is obtained by using a set of simulated runtime SPAF values and the aging model described in Sec. II. For mem_ctrl, Case I is applicable, but for i2c the bound from Case I exceeds the UofM bound. Therefore, Case II is applied to obtain a better bound, as shown in Fig. 4.

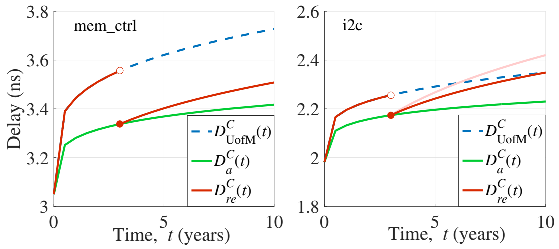


Fig. 4. Example of the recalibration method for aging estimation in two circuits: mem_ctrl follows Case I (left), while i2c follows Case II (right).

B. Post-recalibration aging estimation in the CUT

In this section, we show how the aging estimation scheme in Sec. III-B is modified to incorporate recalibration. At each measurement instant, t_{m_j}, we measure the CUT and the ROSC delays, denoted by $D_a^C(t_{m_j})$ and $D_a^R(t_{m_j})$, respectively. The K_B and K_H values of the ROSC are recalibrated to $K_{j,B}^R$ and $K_{j,H}^R$, respectively, based on the methodology described in Case I of the previous section (Case II does not arise for the ROSC since it has only one path). Similarly, by measuring the circuit, we apply Theorem 3 to determine its $K_{j,B}^C$ and $K_{j,H}^C$ values in the interval $[t_{m_j}, t_{m_{j+1}})$. We then modify Eq. (20) after each t_{m_j} as:

$$\xi_{j,B}^C = \frac{K_{j,B}^C}{K_{j,B}^R} \qquad \xi_{j,H}^C = \frac{K_{j,H}^C}{K_{j,H}^R} \tag{35}$$

where $\xi_{j,B}^C$ and $\xi_{j,H}^C$ are the recalibrated degradation ratios which need to be updated in the LUT after each measurement instant, along with the actual CUT delay, $D_a^C(t_{m_j})$ at t_{m_j}.

The above procedure updates the parameters used for the upper bound in $[t_{m_j}, t_{m_{j+1}})$. Recall that this procedure is to be applied infrequently through the lifetime of the circuit. A more frequent operation is to measure the ROSC only, and to use the parameters of the bound to estimate the CUT delay. We will now explain how this is performed.

Based on the ROSC measurement and the stored ROSC delay at time, t_{m_j}, we first obtain the delay degradation of the ROSC in the interval between t and t_{m_j} as $\Delta D_{B,j}^R(t)$ and $\Delta D_{H,j}^R(t)$ due to BTI and HCI aging, respectively, using the silicon odometer [12]. We then infer $K_{post,B}^R = \frac{\Delta D_{B,j}^R(t)}{\Delta f_j(t)}$ and $K_{post,H}^R = \frac{\Delta D_{H,j}^R(t)}{\Delta g_j(t)}$ from these measurements. We then obtain the CUT delay degradation as:

$$\begin{aligned}\Delta D_{post,j}^C(t) &= K_{post,B}^C \Delta f_j(t) + K_{post,H}^C \Delta g_j(t) \\ &= \xi_{j,B}^C K_{post,B}^R \Delta f_j(t) + \xi_{j,H}^C K_{post,H}^R \Delta g_j(t)\end{aligned} \tag{36}$$

where $K_{post,B}^C$ and $K_{post,H}^C$ correspond to, respectively, the K_B and K_H values of the CUT delay trajectory after each measurement instant, which we infer from the frequent ROSC measurements.

Note that these operations are similar to those in Sec. III-B, except that the degradation is relative to the delay at t_{m_j}, not t_0. Hence the estimated CUT delay with recalibration, $D_{post}^C(t)$, can be obtained by adding $D_a^C(t_{m_j})$, the measured delay at time, t_{m_j}, that is stored in the LUT, to $\Delta D_{post,j}^C(t)$.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

V. Experimental results

The ideas presented in this paper were exercised on a set of representative circuits from the ISCAS'89 [34], ITC'99 [35], and the IWLS'05 [33] benchmark suites. Since the various model parameters and the V_{th} degradation equations are well documented in the public domain for 45nm, the circuits were synthesized using the NanGate 45nm Open Cell Library [36]. Each gate within the library was characterized for nominal delay, output slew and delay sensitivities to V_{th} variation of NMOS and PMOS devices (for both rise and fall transitions) by transistor level HSPICE simulations, and the circuits were synthesized using Synopsys Design Compiler [37]. The simulations were carried out at 125°C and 1.2V. Although we use the RD model for BTI aging in our experiments to show the applicability of the proposed methodology, it is also valid when BTI aging follows a logarithmic function of time, as described by the CT model. The lifetime, t_f, of each CUT has been assumed to be 10 years when both BTI and HCI are significant [26], and we consider $t_0 = 0$.

The difference between various manufactured circuits lies primarily in the variations that they experience due to process and environment effects, and due to the different SPAF values associated with their usage. As explained in Sec. III-C, the impact of process and environment variations is minimal, and hence the primary difference lies in the SPAF values. Some circuits may exercise a CUT frequently and correspond to active SPAFs, while others may be used less often and may correspond to inactive SPAFs. Therefore, we can use the SPAF value as a way to model how a circuit is used in the field. In our experiments, the SP and AF of each gate input of the CUT were assumed to be unity to emulate the worst-case workload. In a real workload, the input SPs are typically biased towards 0 or 1; hence, to emulate such a workload, we generated SPs from a bimodal distribution with peaks at SP = 0.1 and SP = 0.9, in consistence with [38], and set the input AFs to $2s(1-s)$, where s is the SP of that input. To generate a sample of true delay values for the CUT over time, we generated a sample of these input SPs and AFs, propagated them to internal nodes of the circuit, and performed aging-aware STA on it using these SPAF values to simulate the aging of the circuit under an actual workload. The aging-aware STA engine was developed in C++, and the experiments were performed on a 64-bit Ubuntu server with a 3GHz Intel® Core™2 Duo CPU E8400 processor.

While we did not implement the method in silicon, our aging models are based on approaches that are widely accepted in the reliability community, and have been validated by other researchers using experimental silicon measurements.

A. Aging estimation from ROSCs using the UofM bound

We obtain the K_B and K_H values for benchmark circuits, whose names and gate counts, $|G|$, are listed in the first two columns of Table II, by performing aging-aware STA on them under the worst-case workload assumption and using Theorem 1. We also obtain the K_B and K_H values corresponding to 33-stage ROSC sensors [12] using techniques described in Sec. III-A. The degradation ratios, ξ_B^C and ξ_H^C, are listed in the third and fourth columns, respectively, of Table II. The runtime in seconds, τ, required to generate degradation ratios for each circuit is listed in the fifth column. The estimated post-silicon delay, $D_{post}^C(t)$, of the aged CUT is obtained by multiplying the degradation ratios with the delay degradation of the ROSC, and adding the result with the nominal delay, $D_{pre}^C(t_0)$, of the CUT.

The accuracy of our scheme is evaluated by the root mean square error between the CUT delay and the corresponding estimated post-silicon delay over n time instants, $t_j \in [t_0, t_f]$, and is represented by ΔE_{rms}, as:

$$\Delta E_{rms} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{D_{post}^C(t_j) - D^C(t_j)}{D^C(t_j)}\right)^2}, \quad t_j \in [t_0, t_f] \tag{37}$$

where $D^C(t)$ is the true delay of the CUT under worst-case workload, and the error is sampled every half-year interval. The last column in Table II lists ΔE_{rms} for each circuit, expressed in percentage.

TABLE II
Degradation ratios from presilicon analysis.

CUT, C	$\lvert G\rvert$	ξ_B^C	ξ_H^C	τ (s)	ΔE_{rms}
mem_ctrl	6086	1.98	2.35	30	0.001%
wb_dma	2313	1.19	1.61	15	0.096%
ac97_ctrl	8422	1.01	1.49	41	0.003%
i2c	550	1.07	1.28	7	0.028%
aes_core	23104	1.19	1.79	82	0.505%
b15	5581	2.90	3.51	28	0.091%
b17	16531	2.84	3.40	75	0.001%
b20	21625	5.92	9.73	85	0.706%
b21	21661	6.25	8.28	86	0.501%
b22	32513	6.49	8.57	125	0.033%
s5378	692	0.72	0.99	9	0.492%
s13207	594	0.83	0.86	8	0.001%
s15850	340	0.91	0.85	6	0.001%
s38417	4615	1.12	1.86	28	0.558%
s38584	4633	1.21	1.25	26	0.001%

Clearly, the estimated delays match very well with the actual delays, since the ΔE_{rms} values are negligible. Additionally, the modest runtimes for the large benchmark circuits indicate that our method is fast, and hence scalable to real circuits.

Next we study the effect of the (V_{dd}, T)-independence assumptions of the degradation ratios, on the aging estimate as described in Sec. III-C. For this, we use the ξ_B^C and ξ_H^C of each circuit, C, from Table II which were computed at a particular (V_{dd}, T) condition, (1.2V, 125°C), and multiply them with the delay degradation of the ROSC obtained at different (V_{dd}, T) conditions. As before, we add the result to the nominal delay of the CUT to estimate its post-silicon delay, $D_{post}^C(t)$, at these new operating conditions. For a set of representative benchmark circuits, we plot the corresponding estimated delays, $D_{post}^C(t)$, at the end of lifetime, $t_f = 10$ years, along with the actual delay values under the worst-case workload, obtained at several (V_{dd}, T) values in Fig. 5. The estimated delays show an excellent match with the true values

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

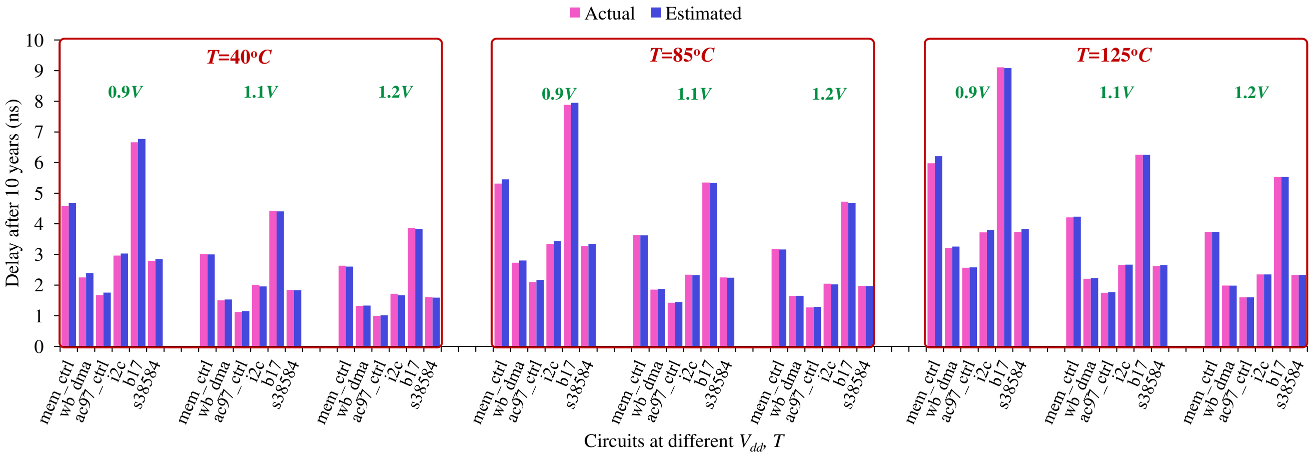


Fig. 5. Estimated delays across different (V_{dd}, T) from degradation ratios at a fixed (V_{dd}, T), indicating V_{dd} and T independence the UofM scheme.

across all nine combinations. Therefore, the degradation ratios are practically independent of V_{dd} and T, and it suffices to compute them at a single (V_{dd}, T) for each circuit.

B. Post-recalibration CUT aging estimation from ROSCs

Although the previous method is very easy to implement and provides a good enough indication of circuit aging with practically no extra design effort and overheads, we can obtain even better aging estimates with some additional circuitry for sensor recalibration involving CUT measurement and updating the LUT, which requires a little more design effort.

1) Speed Wastage Factor: To observe the advantage of recalibrating the sensors, we first define a metric based on the inherent pessimism of the worst-case workload assumption. For this, we perform Monte Carlo simulations on the benchmark circuits with 500 sets of realistic SP and AF values to obtain the temporal trends of their delays under 500 different workloads. Each simulation corresponds to a sample of the realistic input SPs and AFs, propagated throughout a circuit to generate SPs and AFs at internal nodes, and translated into a delay degradation number for each gate. For the i^{th} Monte Carlo run of a circuit, we define its *speed wastage factor* (SWF), or, $\mathit{SWF}(i,t)$, at time, t, expressed in percentage, as:

$$\mathit{SWF}(i,t) = \frac{\mathcal{F}_i(t) - \mathcal{F}_{pre}(t)}{\mathcal{F}_i(t)} \tag{38}$$

where $\mathcal{F}_{pre}(t)$ is the operating frequency set at the presilicon stage with a worst-case workload assumption, and $\mathcal{F}_i(t)$ is the maximum frequency at which the circuit can function correctly at time, t, without any timing violation, corresponding to the workload characterized by the i^{th} Monte Carlo sample. If the exact workload of the CUT was known, it could have been operated at $\mathcal{F}_i(t)$ which is greater than $\mathcal{F}_{pre}(t)$. However, since the workload is unknown, the operating frequency is set to $\mathcal{F}_{pre}(t)$, considering a worst-case aging scenario, so that the CUT is guaranteed to function correctly during its lifetime under aging. The SWF is thus an indication of the performance margin left on the table while assuming the worst-case workload on the CUT, and the *lower* the SWF, the *better* is the performance of the CUT.

To observe the cumulative wastage over the entire lifetime of the CUT, we further define a vector, $\overline{\boldsymbol{SWF}}$, whose i^{th} element, $\overline{\mathit{SWF}}(i)$, is the sampled average of $\mathit{SWF}(i,t)$ at time instants, $t = t_j$, during the CUT lifetime, as:

$$\overline{\mathit{SWF}}(i) = \frac{1}{n}\sum_{j=1}^{n} \mathit{SWF}(i,t_j), \quad t_j \in [t_0, t_f] \tag{39}$$

where $i = 1, \cdots, 500$ corresponds to the 500 simulated workload scenarios.

The mean and range (minimum to maximum) of $\overline{\boldsymbol{SWF}}$ of a set of benchmark circuits without sensor recalibration, is depicted by the first set of bars in Fig. 6. This set of bars correspond to a case where a single measurement is performed at the beginning of lifetime of the CUT, i.e., the operating frequency is determined at the presilicon stage. The average height of this set of bars is 8.78%. In other words, if the aging sensor is calibrated assuming a pessimistic worst-case circuit workload, the circuit is operated at a frequency that is, on average, 8.78% slower than its true capability, consuming unnecessary power/area overheads. We aim to reduce this pessimism using the concept of sensor recalibration by infrequent CUT delay measurements (Sec. IV).

2) Choice of measurement instants: The choice of the measurement instants during the lifetime of a circuit is crucial in determining the extent of pessimism reduction. For various choices of a single measurement instant, t_{m_1}, over the lifetime of the circuit, we plot the statistics of $\overline{\boldsymbol{SWF}}$ over our Monte Carlo simulations for various benchmark circuits in Fig. 6, assuming the same $t_f = 10$ years for all circuits. In other words, we choose to perform recalibration at a single instant, t_{m_1}, through the entire 10-year lifetime, and observe the impact of this choice on the average SWF in Fig. 6. For example, year=0 in the figure corresponds to a (redundant) recalibration at the beginning of life, which means that no recalibration is performed during the lifetime of the CUT, and only presilicon analysis is used to set its operating frequency.

Although $\overline{\boldsymbol{SWF}}$ is reduced for any $t_{m_1} > 0$, there exists a global minimum in $\overline{\boldsymbol{SWF}}$ for a certain choice of the single measurement instant, which is at around 2 years in Fig. 6.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

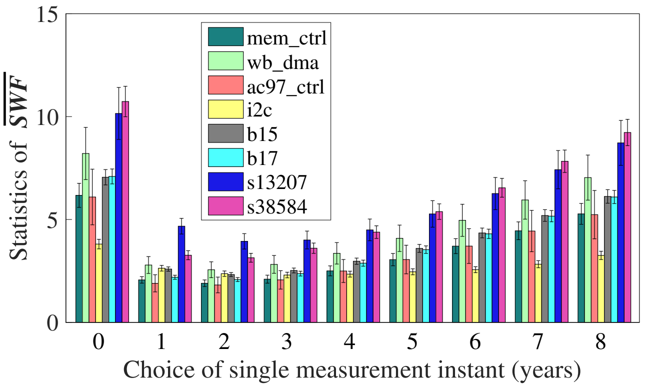


Fig. 6. Effect of the choice of t_{m_1} on average SWF.

We thus attempt to optimize the choice of the N measurement instants which will minimize the SWF. Since we consider circuits with lifetime of 10 years, where both BTI and HCI are prevalent, we heuristically choose the interval between the measurement instants linearly in $f(t)+g(t)$. For N measurement instants in (t_0, t_f), the i^{th} measurement instant, t_{m_i}, is thus obtained by solving the equation:

$$f(t_{m_i}) + g(t_{m_i}) = \left(\frac{i}{N+1}\right)(f(t_f) + g(t_f)) \tag{40}$$

This nonlinear equation does not have a closed-form solution but can easily be solved numerically using Mathematica [39]. For convenience, each solution, t_{m_i}, is rounded off to the nearest half-year.

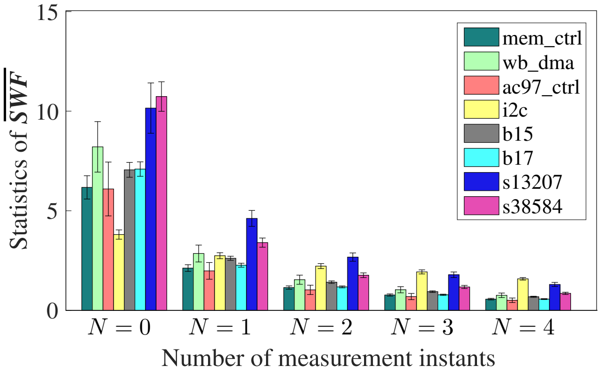


Fig. 7. Reduction in average SWF with number of measurement instants.

Using Eq. (40) to obtain the N measurement instants for $N = 1, 2, 3, 4$, and assuming $t_f = 10$ years for each circuit, we show the mean and range of $\overline{\boldsymbol{SWF}}$ in Fig. 7. The SWF is guaranteed to reduce monotonically in all cases, due to the intermediate true delay measurements of the CUT that refine the aging estimate, and for $N = 4$, the average SWF is less than 1%. In almost all cases, a steep reduction is seen at $N = 1$. The exception is the circuit, i2c, for which the reduction is less than other circuits. This occurs because the recalibration scheme for i2c follows Case II in Theorem 3, i.e., it fits the estimated delay curve between the actual delay at the current measurement instant and the worst-case delay at the next. This case was depicted in Fig. 4 (right) for $N = 1$ under one SPAF sample, corresponding to a specific workload. In general, Case II is seen to rarely arise, except for small circuits such as i2c (with ~ 500 gates): in particular, for large circuits that have tens of thousands of gates, this scenario was not observed.

Finally, N may be chosen depending on the desired amount of SWF reduction for a particular CUT lifetime, t_f. For example, in Fig. 7, the average SWF is already reduced to below 1% for $N = 4$. Hence, increasing N beyond four does not provide significant improvements in the SWF for the conditions assumed here.

We now examine the data from Fig. 7 in greater detail and focus on the error in $\Delta D^C_{post}(t)$ instead of $D^C_{post}(t)$ (which was incorporated in the SWF metric). This error is quantified by the vector, $\boldsymbol{E_\Delta}$, whose i^{th} element, $E_\Delta(i)$, corresponds to the i^{th} Monte Carlo run using a specific workload, averaged over time instants, $t = t_j$, and expressed as:

$$E_\Delta(i) = \frac{1}{n}\sum_{j=1}^{n}\left[\frac{\Delta D^{C,i}_{post}(t_j) - \Delta D^{C,i}_{a}(t_j)}{\Delta D^{C,i}_{a}(t_j)}\right], \quad t_j \in (t_0, t_f] \tag{41}$$

where $\Delta D^{C,i}_{post}(t_j)$ and $\Delta D^{C,i}_{a}(t_j)$ are, respectively, the estimated post-silicon and actual delay degradation of C, from $t = t_0$ under the i^{th} workload.

Table III reports the statistics of $\boldsymbol{E_\Delta}$ for various values of N. For each circuit, we show the mean and standard deviation of $\boldsymbol{E_\Delta}$ for $N = 0, 1, \cdots, 4$. The $N = 0$ case corresponds to the UofM model-based estimate where the sensors were calibrated based on presilicon analysis only. The columns also show the vector of measurement instants, T_M, expressed in

TABLE III
STATISTICS OF $\boldsymbol{E_\Delta}$ WITH SENSOR RECALIBRATION, EXPRESSED AS PERCENTAGE, FOR VARIOUS SETS OF MEASUREMENT INSTANTS, T_M, IN YEARS.

CUT	$N=0$		$N=1$		$N=2$		$N=3$		$N=4$	
	$T_M = [\,]$		$T_M = [1.5]$		$T_M = [0.5, 3.5]$		$T_M = [0.5, 1.5, 5]$		$T_M = [0.5, 1.5, 2.5, 6]$	
	$\mu_{\boldsymbol{E_\Delta}}$	$\sigma_{\boldsymbol{E_\Delta}}$	$\mu_{\boldsymbol{E_\Delta}}$	$\sigma_{\boldsymbol{E_\Delta}}$	$\mu_{\boldsymbol{E_\Delta}}$	$\sigma_{\boldsymbol{E_\Delta}}$	$\mu_{\boldsymbol{E_\Delta}}$	$\sigma_{\boldsymbol{E_\Delta}}$	$\mu_{\boldsymbol{E_\Delta}}$	$\sigma_{\boldsymbol{E_\Delta}}$
mem_ctrl	70.21	10.83	23.91	3.30	12.25	1.56	8.12	1.02	5.89	0.74
wb_dma	89.62	24.23	30.44	7.78	15.53	3.87	10.27	2.55	7.42	1.85
ac97_ctrl	48.43	15.88	15.85	4.89	7.62	2.44	5.04	1.58	3.64	1.14
i2c	42.19	3.69	29.64	2.38	23.60	1.92	20.34	1.58	16.65	1.09
b15	103.51	11.00	37.89	3.34	19.50	1.58	12.71	0.99	9.13	0.69
b17	95.67	9.29	30.56	2.69	14.86	1.18	9.69	0.74	6.96	0.52
s13207	256.25	96.14	110.44	37.52	61.46	20.09	40.19	13.12	28.92	9.44
s38584	341.49	94.36	105.53	28.62	50.68	13.63	33.13	8.89	23.83	6.39

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

years (excluding t_0), for each N based on Eq. (40) with $t_f = 10$ years.

Consistent with prior observations, there is a significant error for the $N = 0$ case, and this error is reduced as N is increased. However, the pessimism in the upper bound is never completely removed, which is desired since this guarantees timing closure throughout the lifetime of the CUT.

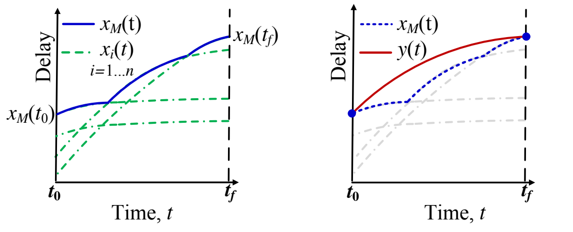
To summarize, our first method based on the UofM bound provides a workable indication of circuit aging with practically no extra effort on part of the designer and a small power/area overhead. A one-time presilicon circuit characterization was used to determine the degradation ratios, and subsequently ROSC measurements were translated to the circuit delay using these ratios. The second proposed method provides a more accurate indication of aging, but requires extra overhead for the intermediate CUT delay measurement and for updating the LUT, and could introduce minor delay overheads ($\sim 1\%$ if the scheme in [20] is employed) in the near-critical paths of the CUT due to the increased fanout load. Although this technique needs higher design effort, the aging estimates obtained are more accurate compared to the first method. However once the test infrastructure is in place, the runtime overhead is negligible for this method since a couple of CUT measurements over 10 years are adequate to bridge the gap in the SWF left by the UofM method.

VI. Conclusion

In this paper we have presented two techniques to estimate aging in circuits due to BTI and HCI, using on-chip ROSC-based sensors. The first one builds upon a presilicon analysis of the CUT, followed by the sensor calibration to translate frequency degradation in the ROSCs to aging in the CUT. During the field operation, post-silicon measurements performed on nearby sensors can be used to estimate aging in the CUT. Since the presilicon analysis is built on the premise of worst-case workload on the CUT, its aging estimate can be further refined using our second proposed technique based on infrequent post-silicon measurements performed directly on the CUT to update the sensor calibration. The updated calibration factors, used in conjunction with the post-silicon measurements on the sensor, can partially capture the real workload of the CUT to yield more accurate aging estimates.

There is a trade-off between the simplicity of implementation versus the accuracy of aging estimates obtained by both techniques. While the first technique is very easy to implement without the need of any additional circuitry other than the ROSC-based sensors, the second one is more accurate, at the cost of higher design effort and overheads due to the additional CUT delay measurement and the LUT update.

Appendix A



(a) (b)

Fig. 8. (a) Maximum among the $x_i(t)$ functions, denoted by $x_M(t)$, (b) Smooth upper-bounded estimate of $x_M(t)$, denoted by $y(t)$.

Proof of Theorem 1: The functions, $x_i(t)$, and their envelope, $x_M(t) = \max_i(x_i(t))$, are depicted in Fig. 8(a).
Since for each $x_i(t)$, we have:

$$x_i(t_f) = x_i(t_0) + \theta_1^i \Delta f(t_f) + \theta_2^i \Delta g(t_f),$$

we obtain the relationship between each θ_1^i and θ_2^i as:

$$\theta_2^i = \frac{\Delta x_i(t_f) - \theta_1^i \Delta f(t_f)}{\Delta g(t_f)} \tag{42}$$

where $\Delta x_i(t) = x_i(t) - x_i(t_0)$.

To ensure that $y(t)$ as depicted in Fig. 8(b) is a tight upper bound on $x_M(t)$, $y(t)$ should satisfy the following conditions:

$$y(t) = x_M(t), \quad t = t_0, t_f \tag{43}$$

$$y(t) \geq x_i(t), \quad \forall t \in [t_0, t_f] \tag{44}$$

Substituting $t = t_f$ in the definition of $y(t)$ in Eq. (13), and using Eq. (43), we obtain:

$$y(t_f) = x_M(t_f) = x_M(t_0) + \theta_1^M \Delta f(t_f) + \theta_2^M \Delta g(t_f) \tag{45}$$

Hence, θ_2^M is computed as:

$$\theta_2^M = \frac{\Delta x_M(t_f) - \theta_1^M \Delta f(t_f)}{\Delta g(t_f)} \tag{46}$$

where $\Delta x_M(t) = x_M(t) - x_M(t_0)$. Using Eqs. (46) and (42), $y(t)$ and each $x_i(t)$ can be rewritten as:

$$y(t) = x_M(t_0) + \theta_1^M \Delta f(t) + \left(\frac{\Delta x_M(t_f) - \theta_1^M \Delta f(t_f)}{\Delta g(t_f)} \right) \Delta g(t)$$

$$x_i(t) = x_i(t_0) + \theta_1^i \Delta f(t) + \left(\frac{\Delta x_i(t_f) - \theta_1^i \Delta f(t_f)}{\Delta g(t_f)} \right) \Delta g(t)$$

Hence the error, $\delta_i(t) = y(t) - x_i(t)$, is obtained as:

$$\begin{aligned} \delta_i(t) &= \Delta x_{M,i}(t_0) \left(1 - \frac{\Delta g(t)}{\Delta g(t_f)} \right) + \Delta x_{M,i}(t_f) \frac{\Delta g(t)}{\Delta g(t_f)} \\ &\quad + (\theta_1^M - \theta_1^i) \Delta f(t_f) \left(\frac{\Delta f(t)}{\Delta f(t_f)} - \frac{\Delta g(t)}{\Delta g(t_f)} \right) \end{aligned} \tag{47}$$

where $\Delta x_{M,i}(t) = x_M(t) - x_i(t)$. Analyzing each addend in Eq. (47) we make the following conclusions:

- Since $x_M(t) = \max_i(x_i(t))$, by definition, both $\Delta x_{M,i}(t_0)$ and $\Delta x_{M,i}(t_f)$ are nonnegative. Since $\Delta g(t) = t^{n_2} - t_0^{n_2}$ with $n_2 \in (0, 1)$, $0 \leq \frac{\Delta g(t)}{\Delta g(t_f)} \leq 1$, $\forall t \in [t_0, t_f]$. Hence the first two addends in Eq. (47) are positive.
- For either form of $f(t) = t^{n_1}$ or $a + b \log t$, $\Delta f(t_f) \geq 0$, and by definition, $\theta_1^M \geq \theta_1^i$ in the third addend.
- Finally, to analyze the last parenthetical expression in Eq. (47) for both $f(t) = t^{n_1}$ and $a + b \log t$, we represent it by $e(t)$ for $t \in [t_0, t_f]$, i.e.,

$$e(t) = \left(\frac{\Delta f(t)}{\Delta f(t_f)} - \frac{\Delta g(t)}{\Delta g(t_f)} \right) \tag{48}$$

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Clearly, $e(t_0) = e(t_f) = 0$, and $e(t)$ is continuous, differentiable and real-valued in $[t_0, t_f]$. Hence by Rolle's Theorem, there exists at least one point, $t = t_1 \in (t_0, t_f)$ for which $e'(t_1) = 0$, where $e'(t)$ is the derivative of $e(t)$ with respect to t. The values of t_1 are obtained as:

$$t_1 = \begin{cases} \left(\frac{n_1}{n_2} \left(\frac{t_f^{n_2} - t_0^{n_2}}{t_f^{n_1} - t_0^{n_1}} \right) \right)^{1/(n_2 - n_1)}, & \text{for } f(t) = t^{n_1} \\ \left(\frac{t_f^{n_2} - t_0^{n_2}}{n_2 \log(t_f/t_0)} \right)^{1/n_2}, & \text{for } f(t) = a + b \log t \end{cases}$$

There is exactly one solution for $e'(t_1) = 0$ for a specific form of $f(t)$, when $n_1,\ n_2 \in (0, 1)$ and $n_2 > n_1$. To observe whether t_1 corresponds to a local maximum or minimum, we further obtain the second derivative of $e(t)$ with respect to t, at t_1, and simplify it as:

$$e''(t_1) = \begin{cases} \frac{-n_1(n_2 - n_1) t_1^{n_1 - 2}}{t_f^{n_1} - t_0^{n_1}}, & \text{for } f(t) = t^{n_1} \\ \frac{-n_2}{t_1^2 \log(t_f/t_0)}, & \text{for } f(t) = a + b \log t \end{cases}$$

Clearly, $e''(t_1) < 0$ for both forms of $f(t)$, implying that t_1 is a maximum, or $e(t)$ is maximum at a single $t = t_1 \in (t_0, t_f)$. Hence, $e(t) \geq 0$ in the entire interval, $[t_0, t_f]$.

Being sum of all positive numbers, $\delta_i(t) \geq 0$ in Eq. (47), $\forall t \in [t_0, t_f]$. Hence $y(t)$ as defined in Eq. (13), is indeed a smooth upper bound for the maximum of $x_1(t), \cdots, x_n(t)$. □

Appendix B

Proof of Theorem 2:

We first compute the K_B value of the UofM-based delay trajectory of C, or K_B^C, as the maximum among the K_B values of all near-critical paths. Let us denote this by $K_B^{p_m}$ for the path, $X = p_m$, of C in Eq. (24). Next we obtain the K_B value of the ROSC delay trajectory, which is simply the K_B value of its single path, r, or K_B^r. Hence using Eq. (24), K_B^C and K_B^R are obtained as:

$$K_B^C = K_B^{p_m} = \mathcal{K}_B^{p_m} F_B(V_{dd}, T) \Delta f(t) \tag{49}$$

$$K_B^R = K_B^r = \mathcal{K}_B^r F_B(V_{dd}, T) \Delta f(t) \tag{50}$$

where $\Delta f(t) = f(t) - f(t_0)$, and $F_B(.)$ is defined in Eq. (26). All terms related to V_{dd} and T in K_B^C and K_B^R in Eqs. (49) and (50) are identical. Hence, for $\xi_B^C = \frac{K_B^C}{K_B^R}$, these terms cancel out in the numerator and denominator, implying that the degradation ratio, ξ_B^C, is independent of V_{dd} and T.

For ξ_H^C, we first obtain the K_H value of the UofM-based delay trajectory of C, or K_H^C using Theorem 1 as:

$$\begin{aligned} K_H^C &= \frac{D^C(t_f) - D^C(t_0) - K_B^C \Delta f(t_f)}{\Delta g(t_f)} \\ &= \frac{D^{p_f}(t_f) - D^{p_0}(t_0) - \Delta D_B^{p_m}(t_f)}{\Delta g(t_f)} \end{aligned} \tag{51}$$

where $\Delta g(t) = g(t) - g(t_0)$, and $D^C(t_f)$ and $D^C(t_0)$ are, respectively, the CUT delays at times, t_f and t_0, when the paths, p_f and p_0, are critical in C. Hence $D^C(t_f)$ and $D^C(t_0)$ have been replaced by $D^{p_f}(t_f)$ and $D^{p_0}(t_0)$, respectively, in the second line of Eq. (51). Similarly, since $K_B^C = K_B^{p_m}$ from Eq. (49) and $K_B^{p_m} \Delta f(t) = \Delta D_B^{p_m}(t)$ from Eq. (24), $K_B^C \Delta f(t_f)$ has been replaced by $\Delta D_B^{p_m}(t_f)$ in Eq. (51). Next we rewrite $D^{p_f}(t_f)$ in Eq. (51) as:

$$D^{p_f}(t_f) = D^{p_f}(t_0) + \Delta D_B^{p_f}(t_f) + \Delta D_H^{p_f}(t_f) \tag{52}$$

where $\Delta D_B^X(t)$ and $\Delta D_H^X(t)$ are the delay shifts for any path, X, due to BTI and HCI, respectively, defined in Eqs. (24) and (25). Hence by substituting this $D^{p_f}(t_f)$ in Eq. (51), we can simplify K_H^C as:

$$K_H^C = \frac{\Delta D^{p_f, p_0}(t_0) + (\Delta D_B^{p_f}(t_f) - \Delta D_B^{p_m}(t_f)) + \Delta D_H^{p_f}(t_f)}{\Delta g(t_f)} \tag{53}$$

where $\Delta D^{p_f, p_0}(t_0) = D^{p_f}(t_0) - D^{p_0}(t_0)$. Next we use the following simplifications:

- $(\Delta D_B^{p_f}(t_f) - \Delta D_B^{p_m}(t_f)) = \Delta \mathcal{K}_B^{p_f, p_m} F_B(V_{dd}, T) \Delta f(t_f)$ using Eq. (24), with $\Delta \mathcal{K}_B^{p_f, p_m} = (\mathcal{K}_B^{p_f} - \mathcal{K}_B^{p_m})$,
- $\Delta D_H^{p_f}(t_f) = \mathcal{K}_H^{p_f} F_H(V_{dd}, T) \Delta g(t_f)$ from Eq. (25), with $F_H(.)$ defined in Eq. (27).

Now to compute ξ_H^C, we need the K_H value of the ROSC delay trajectory, which is simply equal to K_H^r from Eq. (25), where $K_H^r = \mathcal{K}_H^r F_H(V_{dd}, T)$. Hence using this K_H^r, and K_H^C from Eq. (53), we rewrite $\xi_H^C = \frac{K_H^C}{K_H^r}$ as:

$$\xi_H^C = \frac{\Delta D^{p_f, p_0}(t_0) + \Delta \mathcal{K}_B^{p_f, p_m} F_B(V_{dd}, T) \Delta f(t_f)}{\mathcal{K}_H^r F_H(V_{dd}, T) \Delta g(t_f)} + \frac{\mathcal{K}_H^{p_f}}{\mathcal{K}_H^r} \tag{54}$$

This is the expression of ξ_H^C as mentioned in Theorem 2. □

Appendix C

Proof of Theorem 3: We present the proof by mathematical induction in terms of the number, N, of measurement instants.

Let $T_M = \{t_{m_0}, t_{m_1}, \cdots, t_{m_{N-1}}, t_{m_N}\}$ be the set of $N+1$ time instants in $[t_0, t_f]$, where the first N represent the measurement instants, with $t_{m_0} = t_0$ and $t_{m_N} = t_f$ denoting the beginning and end of lifetime of the CUT, C, respectively.

Base Case, $N = 1$:
Here, $T_M = \{t_{m_0}, t_{m_1} = t_f\}$, and the true delay measurement is available at a single instant $t = t_{m_0} = t_0$. Under this condition, $D_{re}^C(t) = D_{\text{UofM}}^C(t)$, which has been proven in Sec. III-A to upper-bound the true circuit delay for $t \in [t_0, t_f]$.

Inductive Hypothesis, $N = r$:
When $T_M = \{t_{m_0}, t_{m_1}, \cdots, t_{m_{r-1}}, t_{m_r}\}$, we assume that $D_{re}^C(t)$, as defined in Eq. (29), together with (30) or (31), provides an upper bound on the true delay for all r measurement instants.

Inductive Step, $N = r + 1$:
Here $T_M = \{t_{m_0}, t_{m_1}, \cdots, t_{m_{r-1}}, t_{m_r}, t_{m_{r+1}}\}$. Using the results of the inductive hypothesis, $D_{re}^C(t)$ forms an upper bound on true delay of the circuit for $t \in [t_0, t_{m_r}]$. At $t = t_{m_r}$, $D_{re}^C(t) = D_a^C(t)$. Now for $t \in [t_{m_r}, t_{m_{r+1}}]$, we consider each of the bounds in Theorem 3.

(I) $D_{re}^{C,I}(t) = D_a^C(t_{m_r}) + K_B^{p_x} \Delta f_r(t) + K_H^{p_x} \Delta g_r(t)$
Hence $D_{re}^{C,I}(t) = D_a^C(t_{m_r}) + (D_{pre}^{p_x}(t) - D_{pre}^{p_x}(t_{m_r}))$, where

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

p_x is a near-critical path of the CUT with maximum delay degradation from t_{m_r} to $t_{m_{r+1}}$ under the worst-case workload characterized at the presilicon stage, among all near-critical paths. Since by definition of $D^{p_x}_{pre}(t)$, no other path under any real workload scenario undergoes as much degradation as p_x, the circuit delay change cannot exceed that of p_x over the interval $t \in [t_{m_r}, t_{m_{r+1}}]$. Hence $D^C_{re}(t)$ forms an upper bound on delay trajectory under any real workload for $t \in [t_{m_r}, t_{m_{r+1}}]$.

(II) $D^{C,II}_{re}(t) = D^C_a(t_{m_r}) + K^{II}_B \Delta f_r(t) + K^{II}_H \Delta g_r(t)$

Now to ensure that $D^C_{re}(t)$ is an upper-bounding curve between t_{m_r} and $t_{m_{r+1}}$ over all paths $p_i \in S_{NC}$, we may simply apply Theorem 1, setting $t_0 = t_{m_r}$, $x_M(t_{m_j}) = D^C_a(t_{m_j})$, and $x_M(t_{m_{j+1}}) = D^C_{pre}(t_{m_{j+1}})$, and this yields the result. □

References

[1] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance CMOS Variability in the 65-nm Regime and Beyond," *IBM J. Res. Dev.*, vol. 50, no. 4.5, pp. 433–449, 2006.

[2] M. A. Alam and S. Mahapatra, "A Comprehensive Model of PMOS NBTI Degradation," *Microelectron. Reliab.*, vol. 45, no. 1, pp. 71–81, 2005.

[3] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "An Analytical Model for Negative Bias Temperature Instability," in *Proc. Int. Conf. Comput.-Aided Design*, 2006, pp. 493–496.

[4] A. Bravaix, C. Guerin, V. Huard, D. Roy, J.-M. Roux, and E. Vincent, "Hot-Carrier Acceleration Factors for Low Power Management in DC-AC Stressed 40nm NMOS Node at High Temperature," in *Proc. Int. Rel. Physics Symp.*, 2009, pp. 531–548.

[5] T. Nigam, "Impact of Transistor Level Degradation on Product Reliability," in *Proc. Custom Integr. Circuits Conf.*, 2009, pp. 431–438.

[6] J. Fang and S. S. Sapatnekar, "Incorporating Hot-Carrier Injection Effects Into Timing Analysis for Large Circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 12, pp. 2738–2751, 2014.

[7] W. Wang, Z. Wei, S. Yang, and Y. Cao, "An Efficient Method to Identify Critical Gates under Circuit Aging," in *Proc. Int. Conf. Comput.-Aided Design*, 2007, pp. 735–740.

[8] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Adaptive Techniques for Overcoming Performance Degradation Due to Aging in CMOS Circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 4, pp. 603–614, 2011.

[9] E. Mintarno, J. Skaf, R. Zheng, J. B. Velamala, Y. Cao, S. Boyd, R. W. Dutton, and S. Mitra, "Self-Tuning for Maximized Lifetime Energy-Efficiency in the Presence of Circuit Aging," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 5, pp. 760–773, 2011.

[10] L. Zhang and R. P. Dick, "Scheduled Voltage Scaling for Increasing Lifetime in the Presence of NBTI," in *Proc. Asian South Pacific-Design Automation Conf.*, 2009, pp. 492–497.

[11] T. H. Kim, R. Persaud, and C. H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 874–880, 2008.

[12] J. Keane, X. Wang, D. Persaud, and C. H. Kim, "An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI, and TDDB," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 817–829, 2010.

[13] K. K. Kim, W. Wang, and K. Choi, "On-Chip Aging Sensor Circuits for Reliable Nanometer MOSFET Digital Circuits," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 57, no. 10, pp. 798–802, 2010.

[14] T. Iizuka, T. Nakura, and K. Asada, "Buffer-Ring-Based All-Digital On-Chip Monitor for PMOS and NMOS Process Variability and Aging Effects," in *Proc. Int. Symp. Design and Diagnostics Electron. Circuits and Syst.*, 2010, pp. 167–172.

[15] T. B. Chan, P. Gupta, A. B. Kahng, and L. Lai, "DDRO: A Novel Performance Monitoring Methodology Based on Design-Dependent Ring Oscillators," in *Proc. Int. Symp. Quality Electron. Design*, 2012, pp. 633–640.

[16] Q. Liu and S. S. Sapatnekar, "Synthesizing a Representative Critical Path for Post-Silicon Delay Prediction," in *Proc. Int. Symp. Physical Design*, 2009, pp. 183–190.

[17] S. Wang, J. Chen, and M. Tehranipoor, "Representative Critical Reliability Paths for Low-Cost and Accurate On-Chip Aging Evaluation," in *Proc. Int. Conf. Comput.-Aided Design*, 2012, pp. 736–741.

[18] D. Sengupta and S. S. Sapatnekar, "Predicting Circuit Aging Using Ring Oscillators," in *Proc. Asian South Pacific-Design Automation Conf.*, 2014, pp. 430–435.

[19] D. Sengupta and S. S. Sapatnekar, "ReSCALE: Recalibrating Sensor Circuits for Aging and Lifetime Estimation under BTI," in *Proc. Int. Conf. Comput.-Aided Design*, 2014, pp. 492–497.

[20] X. Wang, M. Tehranipoor, and R. Datta, "Path-RO: A Novel On-Chip Critical Path Delay Measurement Under Process Variations," in *Proc. Int. Conf. Comput.-Aided Design*, 2008, pp. 640–646.

[21] M. Agarwal, B. C. Paul, M. Zhang, and S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging," in *IEEE VLSI Test Symp.*, 2007, pp. 277–286.

[22] Y. Li, S. Makar, and S. Mitra, "CASP: Concurrent Autonomous Chip Self-Test Using Stored Test Patterns," in *Proc. Design Automation and Test Europe*, 2008, pp. 885–890.

[23] G. I. Wirth, R. D. Silva, and B. Kaczer, "Statistical Model for MOSFET Bias Temperature Instability Component due to Charge Trapping," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2743–2751, 2011.

[24] J.-J. Kim, R. Rao, J. Schaub, A. Ghosh, A. Bansal, K. Zhao, B. Linder, and J. Stathis, "PBTI/NBTI Monitoring Ring Oscillator Circuits with On-Chip Vt Characterization and High Frequency AC Stress Capability," in *Proc. Symp. VLSI Circuits*, 2011, pp. 224–225.

[25] S. Chakravarthi, A. Krishnan, V. Reddy, C. Machala, and S. Krishnan, "A Comprehensive Framework for Predictive Modeling of Negative Bias Temperature Instability," in *Proc. Int. Rel. Physics Symp.*, 2004, pp. 273–282.

[26] C. Schlunder, S. Aresu, G. Georgakos, W. Kanert, H. Reisinger, K. Hofmann, and W. Gustin, "HCI vs. BTI? - Neither One's Out," in *Proc. Int. Rel. Physics Symp.*, 2012, pp. 2F.4.1–2F.4.6.

[27] X. Feng, P. Ren, Z. Ji, R. Wang, K. Sutaria, Y. Cao, and R. Huang, "Novel Voltage Step Stress (VSS) Technique for Fast Lifetime Prediction of Hot Carrier Degradation," in *Proc. Int. Conf. Solid-State and Integr. Circuit Technol.*, 2014, pp. 1–3.

[28] "Reliability PTM model," http://ptm.asu.edu/reliability/.

[29] T. Liu, C.-C. Chen, and L. Milor, "Accurate Standard Cell Characterization and Statistical Timing Analysis using Multivariate Adaptive Regression Splines," in *Proc. Int. Symp. Quality Electron. Design*, 2015, pp. 272–279.

[30] M. Orshansky, L. Milor, L. Nguyen, G. Hill, Y. Peng, and C. Hu, "Intra-field Gate CD Variability and its Impact on Circuit Performance," in *Tech. Dig. Int. Electron Devices Meeting*, 1999, pp. 479–482.

[31] Y. Zheng, A. Basak, and S. Bhunia, "CACI: Dynamic Durrent Analysis towards Robust Recycled Chip Identification," in *Proc. Design Automation Conf.*, 2014, pp. 1–6.

[32] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Predictive Modeling of the NBTI Effect for Reliable Design," in *Proc. Custom Integr. Circuits Conf.*, 2006, pp. 189–192.

[33] "IWLS 2005 Benchmarks," http://iwls.org/iwls2005/.

[34] "ISCAS 1989 Benchmarks," http://www.pld.ttu.ee/~maksim/benchmarks/iscas89/verilog/.

[35] "ITC 1999 Benchmarks," http://www.cad.polito.it/downloads/tools/itc99.html.

[36] "NanGate 45nm Open Cell Library," http://www.si2.org/openeda.si2.org/projects/nangatelib.

[37] Synopsys, Inc., "Design Compiler," http://www.synopsys.com.

[38] D. Lee, D. Blaauw, and D. Sylvester, "Runtime Leakage Minimization through Probability-Aware Dual-V_t or Dual-t_{ox} Assignment," in *Proc. Asian South Pacific-Design Automation Conf.*, 2005, pp. 399–404.

[39] Wolfram Research, Inc., "Mathematica," 2014, Version 10.0.