ReSCALE: Recalibrating Sensor Circuits for Aging and Lifetime Estimation under BTI

Deepashree Sengupta and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455, USA.

Abstract—Bias temperature instability (BTI) induced delay shifts in a circuit depend strongly on its operating environment. While sensors can capture some operating parameters, they are ineffective in measuring vital performance shifts due to changes in the workloads and signal probabilities. This paper determines the delay of an aged circuit by amalgamating more frequent measurements on ring-oscillator sensors with infrequent online delay measurements on a monitored circuit to recalibrate the sensors. Our approach reduces the pessimism in predicting circuit delays, thus permitting lower delay guardbanding overheads compared to conventional methods.

I. Introduction

Bias temperature instability (BTI) is an aging effect that causes the magnitude of the threshold voltage in nanometer-scale circuits to increase under temperature and voltage stress. BTI-induced aging can be partially reversed when the stress is removed, but the general delay trend over long periods shows a degradation over time.

It is crucial to estimate the extent of aging so that remedial techniques can be applied to ensure reliable operation over the lifetime of a circuit. These schemes may be deployed at the presilicon as well as at the post-silicon stage of design. Of necessity, presilicon design must be predicated on the worst-case workload for the circuit so that it is guaranteed to work under all operating conditions. This involves the application of pessimistic guardbands whose power overheads may be excessive and unnecessary for a large fraction of parts in the field. This pessimism could, in principle, be reduced by the use of signal probabilities (SPs) [1] that mimic the operating environment, but this requires foreknowledge of the average workload in real operating conditions in the field, which is often unavailable. Specifically, a circuit that is designed to meet lifetime requirements under a specific average workload may well be used in a very different way by the customer, with very different SP and aging characteristics. As a result, chip design teams often treat SP-based methods with skepticism. Post-silicon techniques, on the other hand, rely on data from surrogate aging sensors [2]–[4], such as ring oscillators, to apply just enough adaptive compensation to mitigate the effect of aging [5], [6]. To a limited extent, they 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. However, aging sensors are surrogates and cannot reflect aging in the circuit with accuracy, since the types of gates in the circuit and the signal stressing patterns and SPs for the two circuits are different.

As an example, let us consider the aging trends of representative circuits of the IWLS 2005 benchmark suite [7] over sets of pseudo-random input probabilities. To reflect signal distributions in real circuits, where the input SPs are typically biased towards 0 or 1 as against the unrealistic "academic" assumption of SP=0.5, these probabilities are generated from a bimodal distribution with peaks at SP=0.1 and SP=0.9 (in consistence with [8]). Each Monte Carlo (MC) simulation corresponds to a sample of these input SPs, propagated throughout a circuit to generate SPs at internal nodes, which are translated into a delay degradation number for each gate. A static timing analysis (STA) run is performed using these degraded gate delays to determine the temporal degradation in the circuit delay. We perform 500 such simulations for each of the circuits.

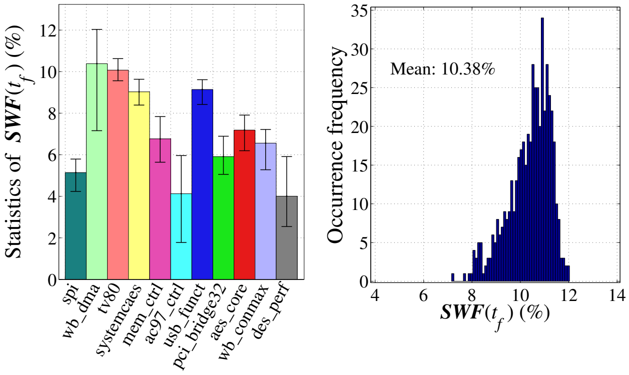

Fig. 1. Effect of worst-case SP assumption on benchmark circuits: left figure depicts amount of pessimism quantified in terms of $\boldsymbol{SWF}(t_f)$ for all the circuits, right figure depicts the histogram of $\boldsymbol{SWF}(t_f)$ for **wb_dma**.

The pessimistic delay for each circuit is obtained by assuming the worst-case workload at every gate input, as in [9]–[11]. Such pessimistic delays, typically used to define a presilicon aging margin, may result in high power/area overheads during optimization [1], [5].

To quantify this pessimism, for each MC run of a circuit, we define the *speed wastage factor*, $SWF(i,t)$, for the i^{th} run as:

$$SWF(i,t) = \frac{f^i(t) - f^{pess}(t)}{f^i(t)} \tag{1}$$

where $f^i(t)$ and $f^{pess}(t)$ are, respectively, the frequencies at time, t, corresponding to the i^{th} MC sample and the worst-case workload case. Since the maximum value of $SWF(i,t)$ occurs at $t = t_f$, we plot the average, minimum and maximum values of the vector, $\boldsymbol{SWF}(t_f)$, whose i^{th} element is $SWF(i,t_f)$, in Fig. 1 (left). For further elaboration, the histogram of the $\boldsymbol{SWF}(t_f)$ for a specific circuit, **wb_dma**, is also shown in Fig. 1 (right) whose mean is 10.38%. In other words, if the aging sensor is calibrated using a pessimistic worst-case probability, this circuit is operated at a frequency about 10% slower than its true capability, consuming unnecessary power/area overheads.

The root cause of the pessimism seen in Fig. 1 is that the prediction is associated with (a) a presilicon characterization and (b) uses a surrogate aging sensor, which has inherent inaccuracies. The worst-case aging trend for a circuit is typically predicted by assuming worst-case stress on all gates. While such a method is guaranteed to be pessimistic, the pessimism may be too large. Some works have attempted to overcome this by using a worst-case stress probability of 0.95 instead of 1.0 on each gate [10], but this is purely empirical. Precise aging information is only obtainable from expensive post-silicon aging measurements performed directly on the circuit [12]–[14] instead of using surrogate sensors. The objective of this work is to build an efficient and precise scheme for diagnosing circuit delay degradation due to aging. Although the scheme can also be extended to any aging mechanisms such as hot carrier injection (HCI) that causes threshold voltages to change over time, our work focuses on BTI, which is the dominant aging mechanism in most products.

Our approach blends the simplicity and low measurement overhead of surrogate sensors with the accuracy of direct measurement. Our scheme avoids the large pessimism gap shown in Fig. 1 by *recalibrating* a set of surrogate sensors based on direct measurements on the circuit to diagnose its actual aging. These measurements are performed only occasionally, thus controlling the high overhead of runtime measurements. Furthermore, we develop a new theory that maps the results of direct measurement to the aging of the surrogate sensor, and propose a framework to recalibrate the surrogate sensors based on measurement data. We demonstrate significant improvement in the speed wastage factor on representative benchmarks and compare our approach with the traditional pessimistic one (Table I).

II. Overview of the Scheme

Our scheme uses an initial calibration of the aging sensor obtained under presilicon worst-case aging of the circuit. Following the runtime measurements on the circuit at a set of *measurement instants*, the sensors are recalibrated to reflect the true past workload of the circuit, and refine its future aging estimate. The aging estimate beyond any measurement instant must assume the worst stressing SPs for the circuit since the data from such instant only provides information about the past and cannot be used to predict its future workload.

Consider a circuit under test (CUT) with a lifetime t_f, and a set of measurement instants, where the i^{th} instant is denoted by t_{m_i}. Let,

- $D_{wc}(t)$ be the worst-case delay curve obtained from presilicon analysis, assuming all gates to be maximally stressed[1].
- $D_{act}(t_{m_i})$ be the measured delay obtained at t_{m_i} corresponding to the actual delay curve of the CUT under its true workload, which is impossible to predict at the presilicon stage. The workload depends on factors such as circuit activity, whether or not the circuit was in sleep mode, and usage statistics.
- $D_{est}(t)$ be the estimated aging curve using our approach. This curve follows $D_{wc}(t)$ until the first measurement instant, and based on $D_{act}(t_{m_i})$, it is recalibrated and modified beyond t_{m_i}.

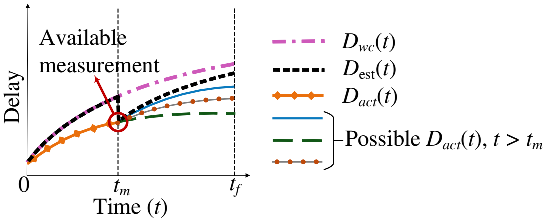

Fig. 2. CUT delay measurement for sensor recalibration: an example.

As a representative outcome of our scheme, consider Fig. 2, where we conceptually depict $D_{wc}(t)$, $D_{act}(t)$ and $D_{est}(t)$ for a single measurement instant, t_m. The multiple curves for $D_{act}(t)$ beyond t_m emphasize the fact that multiple possible future workload scenarios may exist for the CUT based on circuit usage. The $D_{est}(t)$ curve is chosen so that it provides a guaranteed upper bound on the delay beyond t_m by assuming the worst-case workload for the CUT beyond t_m, thus catering to the worst-case scenarios.

To realize our scheme, any functional block must include the aging sensors and the sub-blocks for aging estimation and runtime measurement amidst multiple CUTs as shown in Fig. 3.

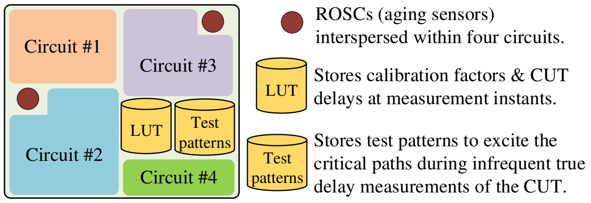
Circuit #1
Circuit #3
Circuit #2
LUT
Test patterns
Circuit #4
ROSCs (aging sensors) interspersed within four circuits.
LUT
Stores calibration factors & CUT delays at measurement instants.
Test patterns
Stores test patterns to excite the critical paths during infrequent true delay measurements of the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

Fig. 3. A functional block equipped with aging estimation tools.

Aging sensors and initial calibration: We use inverter-based ring oscillators (ROSCs) as the aging sensors. Since the aging sensitivities of the inverters in the ROSCs may differ from those of the gates in the CUT, we designate a unique calibration factor, $\mathcal{D}$, that is the ratio of delay degradation of the CUT to that of the ROSC. The value of $\mathcal{D}$ is stored in an on-chip look-up table (LUT) as shown in Fig. 3. This factor is constant between each set of measurement instants.

Between measurement instants, we estimate delay degradation in the CUT by measuring the delay degradation in the ROSC and multiplying it by $\mathcal{D}$. At time zero, we use the worst-case aging curve to obtain $\mathcal{D}$ as in [15]. *This value has been shown to be constant under process, voltage, and temperature variations.*

CUT measurement: At each measurement instant, the true delay of the CUT must be measured and stored in the LUT in Fig. 3. There are several existing schemes that can be used to measure the runtime delay of a circuit such as the Path-RO [13], delay shift circuits [12], [14], or the techniques described in [5]. These schemes require on-chip test patterns (to sensitize critical paths), also shown in Fig. 3.

Sensor recalibration: The aging estimate is based on the initial value of $\mathcal{D}$ assuming worst-case aging of the CUT. However, the actual workload may be different from the worst-case as indicated by the true delay at the measurement instant (e.g. at t_m in Fig. 2). At a small set of such instants, t_{m_i}, the LUT is updated to store $D_{act}(t_{m_i})$. Further, $\mathcal{D}$ is recalibrated and updated, if necessary, in the LUT.

Overhead: The overheads of this scheme are due to (*a*) *Area*: this overhead is low, particularly for large circuits, since a single ROSC can be used for aging prediction in multiple nearby circuits, as quantified in [15]. (*b*) *CUT delay*: the additional load due to the true delay measurement circuitry is negligible and does not significantly affect path delays as demonstrated in [13]. (*c*) *Power*: the switching and leakage power overheads are low since the added circuitry is small. The leakage power of the ROSC circuitry can be further reduced through V_{dd} gating, if necessary. (*d*) *Presilicon analysis*: this is performed just once for each CUT, and its cost is shown to be manageable in Sec. V. (*e*) *Postsilicon recalibration*: This corresponds to the time spent in performing the delay test, ROSC recalibration, and aging estimation. This is performed very infrequently (typically, two or three times during a ten-year chip lifetime), and it can be performed during periods when the CUT is inactive, so that it does not adversely affect the functioning of the chip.

We discuss the initial calibration of the aging sensor in Sec. III-B followed by the measurement and recalibration scheme in Sec. IV.

III. Background

A. Modeling BTI-induced delay degradation

The effect of BTI is to increase the absolute value of the threshold voltage, V_{th}, of both PMOS (by Negative BTI or NBTI) and NMOS (by Positive BTI or PBTI) devices, which causes circuit delays to increase with time. The two most common models used to describe BTI aging are the Reaction-Diffusion (RD) model [16] and the Charge-Trapping (CT) model [17]. Each of these models describes temporal degradation in threshold voltage, $\Delta V_{th}(t)$, as:

$$\Delta V_{th}(t) = c\, h(\xi) f(t) \tag{2}$$

[1] In general, $D_{wc}(t)$ will be a piece-wise differentiable curve if several near-critical paths in the CUT, with different aging sensitivities, become critical over the CUT lifetime. As explained in Sec. III-B, for such a case, we will use a smooth and tight upper-bound on the CUT delay as $D_{wc}(t)$.

where $f(t) \sim t^n$, $n \sim 0.1-0.4$ (by RD model); $f(t) \sim \log t$ (by CT model), $h(.)$ is a function of the stressing SP (which captures BTI recovery effects), ξ, where for PMOS [NMOS], ξ is the probability of signal being low [high], and c depends on the temperature and supply voltage. The function, $f(t)$, can be assumed to be the same for both PMOS and NMOS based on [18]. We note that $f(t)$ is a monotonically increasing function, and we use this property later[2].

The delay of a logic gate undergoing BTI aging is given by $D(t) = D(0) + \mathcal{S}\Delta V_{th}(t)$, where $D(0)$ is the nominal delay of the gate and $\mathcal{S}$ is its delay sensitivity to change in V_{th} computed at the nominal V_{th}. Substituting ΔV_{th} from (2), we obtain:

$$D(t) = D(0) + Kf(t) \tag{3}$$

We refer to the constant, $K = \mathcal{S}\ c\ h(\xi)$, as the K-value of the curve, $D(t)$, which combines the effects of temperature, voltage, and the stressing SP. We observe that under a fixed temperature, supply voltage, and SP, the gate delay increases monotonically with time.

B. Initial ROSC calibration and aging estimation

Fig. 4 depicts an example where a CUT has four near-critical paths, each of whose temporal delay shifts are of the form (3). The paths may have different K-values, and some may become critical during some time period. The delay of the circuit, $D_{CUT}(t)$, is the envelope of these delay curves, and may have points of nondifferentiability where the critical path changes, as shown by the encircled points in the figure. Thus, $D_{CUT}(t)$ is characterized by a set of K-values, one for each differentiable segment of the curve. In contrast, the ROSC has a single critical path with a single K-value, K_{ROSC}, that is typically different from any of the values for the CUT.

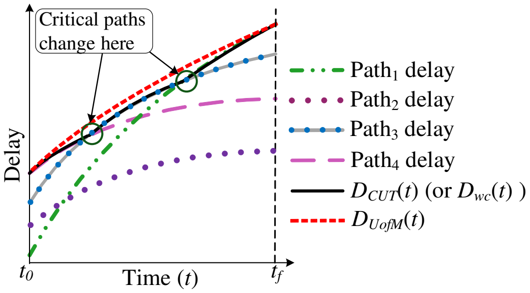

Fig. 4. UofM bound on delay for presilicon worst-case stressing probabilities.

At the presilicon stage, $D_{CUT}(t)$ can be obtained by performing STA on the CUT at multiple points during its lifetime under the worst-case stressing probabilities, hence alternatively called $D_{wc}(t)$ in Fig. 4. A simpler and accurate representation for the CUT delay, which we use in this work, was introduced in [15] using a tight upper-bound on $D_{CUT}(t)$, denoted by $D_{UofM}(t)$ and shown in Fig. 4. For $t \in [t_0, t_f]$, this bound is given by:

$$\begin{aligned} D_{UofM}(t) &= D_{CUT}(t_0) + K_{CUT}(f(t) - f(t_0)) && (4)\\ \text{where } K_{CUT} &= \frac{D_{CUT}(t_f) - D_{CUT}(t_0)}{f(t_f) - f(t_0)} && (5) \end{aligned}$$

This bound is characterized by a single K-value, K_{CUT}, throughout the lifetime of the CUT, and can be computed using just two STA runs. The values of $D_{CUT}(t_0)$ and $D_{CUT}(t_f)$ are obtained by performing STA on the CUT at t_0 and t_f, respectively, assuming the worst-case workload on the CUT. Setting $D_{CUT}(t)$ to this bound, the delay degradations of the CUT and the ROSC at $t \in [t_0, t_f]$ are:

$$\begin{aligned} \Delta D_{CUT}(t) &= K_{CUT}(f(t) - f(t_0)) && (6)\\ \Delta D_{ROSC}(t) &= K_{ROSC}(f(t) - f(t_0)) && (7) \end{aligned}$$

We define a calibration factor, $\mathcal{D}$, for the ROSC as

$$\mathcal{D} = \frac{\Delta D_{CUT}(t)}{\Delta D_{ROSC}(t)} = \frac{K_{CUT}}{K_{ROSC}} \tag{8}$$

The ROSC delay (i.e., the inverse of its oscillating frequency) reflects the delay degradation in the inverters in the ROSC. It is relatively easy to periodically monitor frequency degradation of the ROSC using the concept of *beat frequencies* [2] during runtime. Based on this ROSC measurement, (7) is used to predict the CUT delay from the ROSC delay. By placing the ROSC close enough to the CUT, it may experience similar process variations, and undergo similar temperature and voltage stress, as the CUT. This is shown in [15] to keep the value of $\mathcal{D}$ constant over such variations.

IV. POST-SILICON AGING ESTIMATION

In our scheme, $\mathcal{D}$ and CUT delays at the measurement instants are stored in an LUT, as shown in Fig. 3. The initial $\mathcal{D}$ is obtained from presilicon analysis under the worst-case stress assumption, and is used to translate the delay degradation in the ROSC to that in the CUT until the first measurement instant. At each measurement instant, the CUT delay is measured, providing an accurate view of the actual stressing conditions that it experiences, and $\mathcal{D}$ is recalibrated. The LUT is then updated with the measured delay and the new $\mathcal{D}$. We now explain the theory behind the recalibration procedure, developing a new result in Theorem 1 that generalizes the presilicon upper bound of [15] to our postsilicon approach based on measurement and recalibration.

The input to the procedure is a set, T_M, of $N+1$ time instants, $\{t_0, t_{m_0}, \cdots, t_{m_{N-1}}, t_N\}$, where t_{m_i} is the i^{th} measurement instant, $i = 1 \cdots (N-1)$, $t_0 = 0$, and $t_N = t_f$, and is known before manufacturing. In addition, presilicon analysis provides the worst-case delay of the CUT, $D_{wc}(t)$, at these measurement instants, i.e., $D_{wc}(t_{m_i})$, by performing $(N+1)$ STA runs. We can also use $D_{UofM}(t_{m_i})$ instead of $D_{wc}(t_{m_i})$, which would require only two STA runs as explained in Sec. III-B, and reduce presilicon computation at the cost of accuracy[3]. The knowledge of the near-critical paths of the CUT under the worst-case workload is also available from the presilicon analysis.

A. CUT delay estimate post-measurement

We define a factor, K_f^m, as the maximum among the K-values (Sec. III-B) of the near-critical paths under the worst-case aging, i.e.,

$$K_f^m = \max_{i \in S_{NC}} \left[\frac{D_{wc}^i(t_f) - D_{wc}^i(t_0)}{f(t_f) - f(t_0)} \right] \tag{9}$$

where, S_{NC} is the set of near-critical paths, and $D_{wc}^i(t_f)$ and $D_{wc}^i(t_0)$ are the delays of the i^{th} such path at t_0 and t_f, respectively. In addition to the runtime delay measurement, we use K_f^m to estimate delay of the CUT. The rationale behind using K_f^m is that it is the largest among K-values of the near-critical paths under both the worst-case and realistic workload. Therefore, if the post-measurement delay estimate upper-bounds any path with K-value less than or equal to K_f^m, it is guaranteed to cater to the absolute worst-case future workload of the CUT. The estimated delay $D_{est}(t)$, is obtained based on Theorem 1 (proof deferred to the Appendix).

[2] BTI exhibits partial recovery during which V_{th} is slightly reduced. Equation (2) represents the envelope of degradation, which increases monotonically.

[3] Loss of accuracy is minimal as $D_{UofM}(t)$ is a tight upper-bound [15].

Theorem 1 *Let T_M be indexed by $c = 0, 1, \cdots, N$, and for $t \in [T_M(c), T_M(c+1)]$, let the estimated delay be given by:*

$$D_{est}(t) = D_{est}(T_M(c)) + K_{min}(f(t) - f(T_M(c))) \tag{10}$$

where K_{min} is obtained as:

$$K_{min} = \min\left(K_f^m, \left(\frac{D_{wc}(T_M(c+1)) - D_{act}(T_M(c))}{f(T_M(c+1)) - f(T_M(c))}\right)\right) \tag{11}$$

The above equations are executed recursively at every $T_M(c)$, from $c = 0$ to $N-1$. Then $D_{est}(t)$ is an upper-bound on the actual delay of the CUT under every possible realistic workload.

Theorem 1 provides a single K-value, K_{min}, of the aging curve of the CUT between each set of consecutive measurement instants. Referring to Fig. 2, K_{min} is the the K-value of the delay estimate, $D_{est}(t)$, beyond t_m, based on the measured CUT delay, $D_{act}(t_m)$.

The first argument of the min function in (11) recalibrates the K-value. In theory, it is possible for this recalibrated equation to exceed the $D_{UofM}(t)$ bound at some time instants, although this does not happen in our experiments. The second argument of the min function ensures that K_{min} can never exceed the $D_{UofM}(t)$ bound.

Fig. 5 shows an application of Theorem 1 with three random measurement instants for **wb_dma**. We simulate a realistic workload as explained in Sec. I to obtain $D_{act}(t_{m_i})$, $i = 1, \cdots, 3$. In this circuit, both the UofM bound and pessimistic trajectory, $D_{wc}(t)$ are the same due to a dominant critical path. Hence the estimated delay, $D_{est}(t)$ follows $D_{wc}(t)$, until t_{m_1}. Beyond t_{m_1}, $D_{est}(t)$ is updated based on the true measured delays. The dotted lines show the trajectories $D_{est}(t)$ would have followed if further measurement and recalibration were not performed.

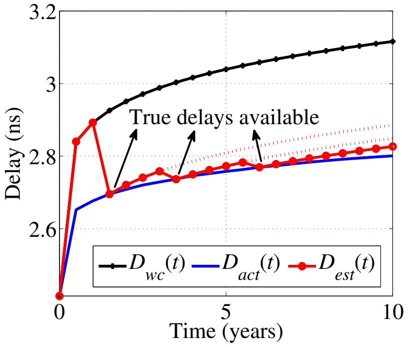

Fig. 5. Use of Theorem 1 on **wb_dma** with three measurement instants.

The estimated delay thus incorporates past workload information to refine future estimate of the delay.

B. Sensor recalibration and aging estimation

Based on the above discussion, it can be seen that after the i^{th} measurement instant, t_{m_i}, the K-value of the aging curve of the CUT, K_{min}, is described by (11). The delay degradations of the CUT and the ROSC from t_{m_i} are therefore expressed as:

$$\Delta D_{CUT}(t) = K_{min}(f(t) - f(t_{m_i})) \tag{12}$$

$$\Delta D_{ROSC}(t) = K_{ROSC}(f(t) - f(t_{m_i})) \tag{13}$$

The calibration factor, $\mathcal{D}$, in (8) for the ROSC is updated to $\mathcal{D}'$ as:

$$\mathcal{D}' = \frac{K_{min}}{K_{ROSC}} \tag{14}$$

The $\mathcal{D}'$ and $D_{CUT}(t_{m_i})$ replace their corresponding old values from the previous measurement instant, stored in the LUT (Fig. 3). The delay degradation of the CUT, $\Delta D_{CUT}(t)$ for $t > t_{m_i}$ is then obtained from the stored $D_{CUT}(t_{m_i})$ and $\Delta D_{ROSC}(t)$ measurements during this period, as explained in Sec. III-B, using $\mathcal{D}'$.

C. Effect of PVT variations

For correct functioning of the proposed framework, the sensors should try to match the process parameter variations, V_{dd}, temperature, and the signal stress probabilities in the CUT. Due to spatial proximity of the ROSCs and the CUT, they face similar temperature stress, and systematic variations within the CUT in any manufactured part are similar to those in the ROSCs close to it. This proximity also enables the ROSCs to connect to the supply lines of the CUT thus capturing the effects of V_{dd} variations under DVFS and power gating.

These factors are incorporated automatically during presilicon analysis (refer Sec. III-B and proof of robustness to PVT variations in [15]). However, the random variations and the effect of stressing SPs (shown to be significant in Fig. 1) in the CUT are difficult to match with those of the ROSC from the presilicon analysis alone. Our approach captures these variations through direct measurement on the CUT at the measurement instants and use of $\mathcal{D}$. In particular, since the calculation of K_{min} integrates the effects of all near-critical paths (which could be potential critical paths in different manufactured parts), our bound in Theorem 1 is robust to all variations.

V. Experimental Setup and Results

A. Experimental setup

The ideas in this paper are exercised on a set of representative IWLS 2005 benchmarks. The circuits are synthesized in Synopsys Design Compiler using the NanGate 45nm Open Cell Library barring the XOR, XNOR and specialized gates (half and full-adders, fill-cells, antennae, tristate gates, and multiplexors). The circuits are aggressively optimized for timing during synthesis. We present our analysis based on the RD model of BTI, and select the value of c in (2) such that there is 25% degradation in PMOS V_{th} (due to NBTI) in 10 years [11]. The degradation in NMOS V_{th} due to PBTI is assumed to be one-third of that due to PMOS NBTI [11].

We perform the simulations at the *typical* process and the *worst-slow* temperature and voltage corner as specified by the NanGate library, i.e., at $T = 125^oC$ and $V_{dd} = 0.9V$. This choice is not critical as our method is robust to variations as explained in Sec. IV-C. The beginning (t_0) and end of lifetime (t_f) are zero and 10 years, respectively. The presilicon analysis of each CUT requires two STA runs (at t_0 and t_f), the runtime of which is less than a minute even for the largest benchmark circuit. To simulate the long-term realistic operating environment of the CUT, we choose a distribution of signal probabilities as in the description of Fig. 1.

B. Choice of N and t_m

Since BTI-induced aging is proportional to a sub-linear function of time, $f(t)$, the maximum degradation occurs towards the beginning of lifetime. Hence the interval between the measurement instants should be chosen linearly in $f(t)$ for the best post-silicon aging estimation.

Under the RD model assumed here, for N measurement instants, we chose the i^{th} measurement instant, t_{m_i}, $i = 1, \cdots, N$ as:

$$t_{m_i} = \left(\frac{i}{M+1}\right)^{1/n} t_f \tag{15}$$

where M is increased from one, until the first N non-negligible values of t_{m_i} are obtained. Each t_{m_i} is rounded off to the nearest half-year for convenience. We present the results for $N = 1, 2, 3, 4$, which correspond to the four sets of t_{m_i} values.

C. Effects of recalibration

We begin with a single circuit, **wb_dma** for recalibration at a single measurement instant, t_m. Using five sets of SPs from $t \in [0, t_m]$, we obtain five realistic aging curves for the CUT until t_m. For each such curve, we apply ten different sets of SPs from t_m to t_f, to obtain a total of 50 realistic workload scenarios. If i indexes the five SP sets from 0 to t_m, and j indexes the ten SP sets from t_m to t_f for each i^{th} set, we obtain five groups of ten actual delay curves.

We define the average speed wastage factor, $\overline{SWF}(k)$ for a particular type of workload (indexed by k), as the time-average of $SWF(k,t)$ defined in (1). We plot $\overline{SWF}(k)$ for the above 50 realistic workload scenarios (i.e., $k = 1, \cdots, 50$) vs. the location of t_m in Fig 6. For example, the first set of ten bars for $t_m = 0$ represents the ten workloads corresponding to $i = 1$ and $j = 1, \cdots, 10$, the second set represents the next ten workloads for $i = 2$ and $j = 1, \cdots, 10$ and so on. Clearly, the goodness of estimation increases initially as t_m is increased as shown by the reduced $\overline{SWF}(k)$, but the advantage of the recalibration scheme slowly reduces as we keep shifting t_m towards t_f. In fact, $\overline{SWF}(k)$ is reduced uniformly over all k irrespective of the actual workload of the CUT when $t_m \sim 1$ year (our choice of t_m for $N = 1$ in (V-B) is consistent with this observation), whereas for $t_m = 0$ or close to t_f, reduction in $\overline{SWF}(k)$ depends on the workload. This is attributed to the fact that BTI is a front-loaded process, and if the aging trend in the CUT is captured early in time (see Sec. V-B), its future aging trend can also be optimally captured by our approach, irrespective of the true workload.

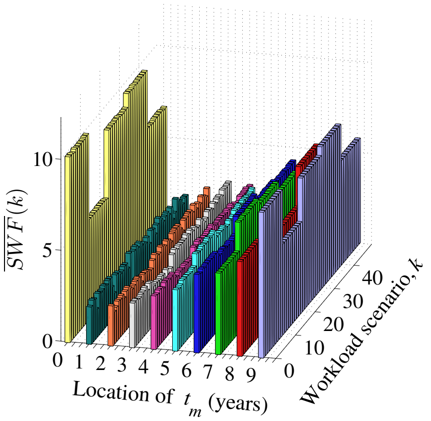

Fig. 6. Average speed wastage, $\overline{SWF}(k)$, vs. temporal shift in t_m in **wb_dma**.

Next, we use the MC simulations as described in Sec. I to obtain the statistics of the $\boldsymbol{SWF}(t_f)$ (defined after (1)) for multiple CUTs with N measurement instants whose locations are chosen as (15). Without any post-silicon measurement, each element of $\boldsymbol{SWF}(t_f)$ is large as shown by their mean and maximum for the $N = 0$ case[4] in Fig. 7 as this is based on the worst-case aging scenarios. Both mean and range of $\boldsymbol{SWF}(t_f)$ decrease with increasing N as the real aging trends of the CUT are captured by the true delay measurements.

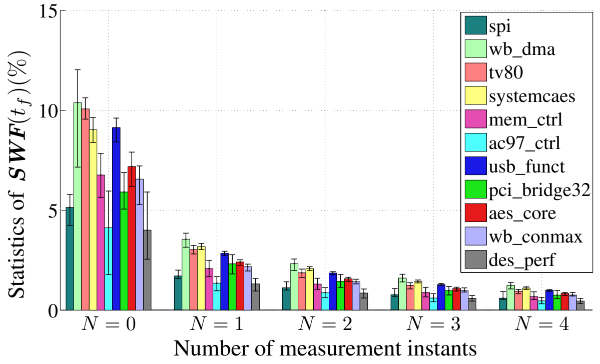

Fig. 7. Range of $\boldsymbol{SWF}(t_f)$ about its mean for N measurement instants.

Let us now look at the data from Fig. 7 in more details and focus on the error in $\Delta D_{est}(t)$ instead of $D_{est}(t)$ (which was incorporated in $SWF(k,t)$). We quantify the error in $\Delta D_{est}(t)$ by the vector, E, whose i^{th} element, corresponds to the i^{th} MC run, and defined as:

$$E(i) = \text{Mean}_t \left[\frac{\Delta D^i_{est}(t) - \Delta D^i_{act}(t)}{\Delta D^i_{act}(t)} \right], \; t > 0 \qquad (16)$$

where $\Delta D^i_{est}(t)$ and $\Delta D^i_{act}(t)$ are the estimated and actual delay degradation of the CUT from $t = 0$, respectively, at the i^{th} run, and the parenthetical expression in (16) is averaged over t to obtain $E(i)$. Table I reports the statistics of E in terms of its mean and range (minimum and maximum) with N. The first column denotes the CUT. The second to fifth columns are each divided into three subcolumns representing the minimum, average, and maximum values of E for $N = 0, 1, \cdots, 4$, respectively. The columns also show the vector of measurement instants, T_M (defined in Sec. IV) , in years, excluding t_0 and t_f in years, for each N based on (15).

As observed in Fig. 5, there is large difference between the worst-case delay curve and the actual delay. This difference varies from circuit to circuit, based on the topology, depth and structure of the near-critical paths. Hence the error in $\Delta D_{est}(t)$ is very high without any post-silicon calibration ($N = 0$ case) as seen in Table I. This error is reduced as N is increased (as seen in both Fig. 5 and Table I), since the true delay measurement of the CUT incorporates information about its past aging scenario to bring down the pessimism. However, pessimism still exists due to the worst-case SP approximation to derive the K-values beyond each measurement instant, because of which the error is always non-zero even after increasing N.

We have also observed that by selecting the measurement instants as (15), we achieve the best results. Any other schedule of selecting these potentially increases the error even if N is increased.

D. Contents of the LUT

The circuits fall in two categories: ones where the multiple critical paths "cross over" as shown in Fig. 4, and others in which the path critical at t_0 remains critical through t_f, among a set of near-critical paths. For the CUTs in first category, the $\mathcal{D}$ values are updated by our proposed methodology, whereas for the second, they remain fairly constant after every measurement instant. However, according to [15], for the current library, the amount of crossover between paths as shown in Fig. 4 is minimal, and the upper-bounded curve is very close to the maximum delay curve. We observed a similar case and most of the CUTs in Table I belonged to the second category in spite of our aggressive timing optimization. Hence, after each measurement instant, barely any change was observed in $\mathcal{D}$ values of the CUTs, and the LUT was updated only with the measured CUT delays.

VI. Conclusion

We have proposed an algorithm to reduce pessimism in estimating BTI-induced aging in digital circuits in terms of their delay degradation. Our estimation scheme is facilitated by both post-silicon runtime measurements and updates to the sensor calibration factor.

Appendix I

In this section, we present a proof of Theorem 1. We begin by presenting a lemma [15] that provides a tight upper-bound on the maximum of a set of monotonically increasing curves:

Lemma 1: *In the interval* $[t_0 \; t_f]$*, an upper-bound on the maximum of a set of monotonically increasing functions* $x_1(t), x_2(t), \cdots, x_{n+1}(t)$ *such that* $x_i(t) = x_i(t_0) + k_i(f(t) - f(t_0))$*, is given by:*

$$y_n(t) = x_M(t_0) + \left[\frac{x_M(t_f) - x_M(t_0)}{f(t_f) - f(t_0)} \right] (f(t) - f(t_0)) \qquad (17)$$

[4] This case is equivalent to the left half of Fig. 1.

TABLE I

ESTIMATION ERROR (PERCENTAGE) IN DELAY DEGRADATION WITH THEIR MINIMUM, AVERAGE AND MAXIMUM VALUES.

CUT	$N = 0$			$N = 1$			$N = 2$			$N = 3$			$N = 4$		
	$T_M = [\,]$			$T_M = [1]$ yr.			$T_M = [0.5, 2.5]$ yrs.			$T_M = [0.5, 1.5, 4]$ yrs.			$T_M = [0.5, 1, 2.5, 5]$ yrs.		
	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.
spi	28.00	38.22	46.78	8.61	9.97	12.05	4.27	4.87	5.99	2.64	2.98	3.80	1.88	2.12	2.82
wb_dma	39.60	82.81	113.15	13.56	21.39	27.15	7.63	10.35	12.56	4.64	6.32	7.82	3.30	4.49	5.49
tv80	135.78	155.93	181.79	32.65	37.25	43.29	14.22	16.61	18.98	8.62	9.90	11.66	6.00	6.97	7.90
systemcaes	78.54	92.05	107.13	21.82	24.82	27.67	11.02	12.19	13.40	6.75	7.42	8.07	4.79	5.27	5.79
mem_ctrl	49.06	67.64	90.90	12.00	16.58	21.56	5.74	7.69	9.59	3.56	4.62	6.08	2.51	3.27	4.33
ac97_ctrl	5.10	16.38	27.15	2.57	4.24	5.80	1.41	2.05	2.70	0.77	1.25	1.67	0.58	0.89	1.18
usb_funct	86.90	105.30	118.41	22.07	25.49	28.17	10.75	11.92	13.18	6.43	7.27	7.89	4.63	5.17	5.67
pci_bridge32	34.43	44.41	58.05	10.00	13.70	17.16	4.96	6.79	8.84	3.07	4.09	5.20	2.21	2.86	3.66
aes_core	58.14	78.79	96.21	16.85	20.49	23.71	8.29	9.96	11.32	5.38	6.09	6.91	3.77	4.33	4.90
wb_conmax	36.50	53.04	62.87	11.14	13.71	15.70	5.83	6.65	7.48	3.55	4.06	4.63	2.52	2.89	3.34
des_perf	10.37	19.12	32.78	3.09	4.95	7.36	1.54	2.41	3.12	1.06	1.47	1.90	0.70	1.05	1.35

where the function $x_M(t) = \max_{i \in 1 \cdots n+1}(x_i(t))$ represents the upper envelope of the functions x_1 through x_{n+1}.

Referring to Fig. 4 and (3), the path delays can be expressed like $x_i(t)$. The maximum of the path delays is the delay of the CUT represented by the piece-wise smooth curve $D_{CUT}(t)$ in the Fig. 4. Lemma 1 provides the upper-bound, $D_{UofM}(t)$, on $D_{CUT}(t)$ derived in a similar fashion as the $y_n(t)$ for the maximum of the $x_i(t)$'s.

Proof of Theorem 1: Based on (3), delay of any near critical path, p_i, under realistic workload, can be expressed as $D^a_{p_i}(t) = D^a_{p_i}(0) + K^a_{p_i}(t)$, such that the actual delay of the CUT is the maximum over such path delays, i.e., $D_{act}(t) = \max_{p_i}[D^a_{p_i}(t)]$. Similarly, delay of the same path under the worst-case workload can be expressed as $D^m_{p_i}(t)$, such that $D_{wc}(t) = \max_{p_i}[D^m_{pi}(t)]$. This is represented by $D_{CUT}(t)$ in Fig. 4.

From Lemma 1, we only need $D_{wc}(0)$ and $D_{wc}(t_f)$ to obtain the upper-bounded delay curve with a single K-value. Since there is no aging at $t = 0$, $D_{wc}(0) = D_{act}(0)$, which also implies that $D_{wc}(0) = \max_{p_i}[D^a_{p_i}(0)]$. Similarly, the worst-case delay is always more than the actual delay because of which $D_{wc}(t_f) \geq \max_{p_i}[D^a_{pi}(t_f)]$.

Keeping the above conclusions in mind, we now begin the proof formally. We use mathematical induction on the number of measurement instants, N, to show that $D_{est}(t) \geq D^a_{p_i}(t)$, $\forall p_i$ and $t \in [0, t_f]$.

Basis case: For $N = 1$, $T_M = \{0, t_f\}$ and $D_{act}(0) = D_{wc}(0)$. Since by definition of K^m_f, $D_{wc}(0) + K^m_f f(t_f) \geq D_{wc}(t_f)$, K_{min} in (11) is $\left\lceil \frac{D_{wc}(t_f) - D_{wc}(0)}{f(t_f)} \right\rceil$. The $D_{est}(t)$ is obtained using (10) as:

$$D_{est}(t) = D_{wc}(0) + \left(\frac{D_{wc}(t_f) - D_{wc}(0)}{f(t_f)} \right) f(t) \qquad (18)$$

From Lemma 1, $D_{est}(t)$ forms an upper-bound on the maximum of path delays under worst-case workload. Since the realistic workload is always more relaxed than the worst-case one, $D_{est}(t) \geq D^a_{p_i}(t)$, $\forall p_i$ and $t \in [0, t_f]$.

Inductive hypothesis: For $N = r$, $T_M = \{0, t_{m_1} \cdots t_{m_{r-1}}, t_f\}$. Let $D_{est}(t)$ as defined in (10) form an upper-bound on the CUT delay for any realistic workload. In other words, $D_{est}(t) \geq D^a_{p_i}(t)$, $\forall p_i$ and $\forall t \in [0, t_f]$.

Inductive step: For $N = r + 1$, $T_M = \{0, t_{m_1} \cdots t_{m_{r-1}}, t_{m_r}, t_f\}$, and we can ascertain that $D_{est}(t)$ as defined in (10) forms an upper-bound on $D^a_{p_i}(t)$, $\forall p_i$ and $t \in [0, t_{m_r}]$ from the inductive hypothesis.

Now for $t \in (t_{m_r}, t_f]$, depending on $D_{act}(t_{m_r})$ and K^m_f, K_{min} is either K^m_f **(Case-1)** or $\left(\frac{D_{wc}(t_f) - D_{act}(t_{m_r})}{f(t_f) - f(t_{m_r})} \right)$ **(Case-2)**.

Case 1: Let $e_{1i}(t) = D_{est}(t) - D_{p_i}(t)$. Then,

$$e_{1i}(t) = \left(D_{act}(t_{m_r}) - D^a_{p_i}(t_{m_r})\right) + (K^m_f - K^a_{p_i})(f(t) - f(t_{m_r})) \quad (19)$$

Each of the parenthetical expressions in $e_{1i}(t)$ is positive due to the following:

1. Due to monotonic property of $f(t)$, $f(t) > f(t_{m_r})$ for $t \in (t_{m_r}, t_f]$.

2. By definition, $K^m_f \geq K^m_{p_i}$.

3. The actual delay $D_{act}(t_{m_r}) \geq D^a_{p_i}(t_{m_r})$, since $D_{act}(t_{m_r})$ was obtained on account of having maximum delay among all the near-critical paths.

Being the sum of all positive numbers, $e_{1i}(t) \geq 0$, $\forall p_i$ and $t \in (t_{m_r}, t_f]$.

Case 2: Let $e_{2i}(t) = D_{est}(t) - D_{p_i}(t)$. Then,

$$\begin{aligned} e_{2i}(t) = D_{act}(t_{m_r}) - D^a_{p_i}(t_{m_r}) + \\ \left(\frac{D_{wc}(t_f) - D_{act}(t_{m_r})}{f(t_f) - f(t_{m_r})} - K^a_{p_i} \right) (f(t) - f(t_{m_r})) \end{aligned} \qquad (20)$$

Since $D_{est}(t)$ and $D_{p_i}(t)$ increase monotonically, $e_{2i}(t)$ either monotonically increases or decreases in $t \in (t_{m_r}, t_f]$. By algebraic manipulation, $e_{2i}(t_{m_r}) = D_{act}(t_{m_r}) - D^a_{p_i}(t_{m_r}) \geq 0$, and $e_{2i}(t_f) = D_{wc}(t_f) - D^a_{p_i}(t_f) \geq 0$ (by definition). Hence $e_{2i}(t) \geq 0$, $\forall p_i$ and $t \in (t_{m_r}, t_f]$.

For both cases, $D_{est}(t) \geq D^a_{p_i}(t)$, $\forall p_i$ and $t \in (t_{m_r}, t_f]$. Hence $D_{est}(t) \geq D^a_{p_i}(t)$, $\forall p_i$ and $t \in [0, t_f]$ implying that $D_{est}(t)$ still forms an upper-bound on the maximum of the path delays. Thus $D_{est}(t)$ as given by (10) indeed forms an upper-bound on CUT delay $\forall t \in [0, t_f]$. □

ACKNOWLEDGMENT

This work was supported in part by the SRC award 2012-TJ-2234 and the NSF (grant CCF-1017778, CCF-1162267).

REFERENCES

[1] S. V. Kumar, *et al.*, "NBTI-aware synthesis of digital circuits," in *Proc. DAC*, pp. 370–375, 2007.

[2] T. H. Kim, *et al.*, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE J Solid-St. Circ.*, vol. 43, pp. 874–880, April 2008.

[3] T. B. Chan, *et al.*, "DDRO: A novel performance monitoring methodology based on design-dependent ring oscillators," in *Proc. ISQED*, pp. 633–640, 2012.

[4] S. Wang, *et al.*, "Representative critical reliability paths for low-cost and accurate on-chip aging evaluation," in *Proc. ICCAD*, pp. 736–741, 2012.

[5] E. Mintarno, *et al.*, "Self-tuning for maximized lifetime energy-efficiency in the presence of circuit aging," *IEEE T Comput. Aid. D.*, vol. 30, pp. 760–773, May 2011.

[6] H. Mostafa, *et al.*, "NBTI and process variations compensation circuits using adaptive body bias," *IEEE T Semiconduct. M.*, vol. 25, pp. 460–467, August 2012.

[7] "IWLS 2005 Benchmarks." `http://iwls.org/iwls2005/`.

[8] D. Lee, *et al.*, "Runtime leakage minimization through probability-aware dual-V$_t$ or dual-t$_{ox}$ assignment," in *Proc. ASP-DAC*, pp. 399–404, 2005.

[9] W. Wang, *et al.*, "Statistical prediction of circuit aging under process variations," in *Proc. CICC*, pp. 13–16, 2008.

[10] M. Agarwal, *et al.*, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. ITC*, pp. 1–10, 2008.

[11] S. V. Kumar, *et al.*, "Adaptive techniques for overcoming performance degradation due to aging in CMOS circuits," *IEEE T VLSI Syst*, vol. 19, pp. 603–614, April 2011.

[12] M. Agarwal, *et al.*, "Circuit failure prediction and its application to transistor aging," in *IEEE VLSI Test Symp.*, pp. 277–286, 2007.

[13] X. Wang, *et al.*, "Path-RO: a novel on-chip critical path delay measurement under process variations," in *Proc. ICCAD*, pp. 640–646, 2008.

[14] Y. Li, *et al.*, "CASP: concurrent autonomous chip self-test using stored test patterns," in *Proc. DATE*, pp. 885–890, 2008.

[15] D. Sengupta and S. S. Sapatnekar, "Predicting circuit aging using ring oscillators," in *Proc. ASP-DAC*, pp. 430–435, 2014.

[16] S. Chakravarthi, *et al.*, "A comprehensive framework for predictive modeling of negative bias temperature instability," in *Proc. IRPS*, pp. 273–282, 2004.

[17] R. Da Silva and G. I. Wirth, "Logarithmic behavior of the degradation dynamics of metal oxide semiconductor devices," *J Stat. Mech.-Theory E.*, vol. P04025, pp. 1–12, April 2010.

[18] J. J. Kim, *et al.*, "PBTI/NBTI monitoring ring oscillator circuits with on-chip Vt characterization and high frequency AC stress capability," in *Proc. VLSIC*, pp. 224–225, 2011.