Predicting Circuit Aging Using Ring Oscillators

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

Abstract—This paper presents a method for inferring circuit delay shifts due to bias temperature instability using ring oscillator (ROSC) sensors. This procedure is based on presilicon analysis, postsilicon ROSC measurements, a new aging analysis model called the Upperbound on f_{Max} (UofM), and a look-up table that stores a precomputed *degradation ratio* that translates delay shifts in the ROSC to those in the circuits. This method not only yields delay estimates within 0.2% of the true values with very low runtime, but is also independent of temperature and supply voltage variations.

I. Introduction

Bias Temperature Instability (BTI) is a pressing reliability issue that degrades the threshold voltages (V_{th}) of nanometer-scale MOS devices during normal circuit operation under voltage and temperature stress. The degradation in PMOS [NMOS] is called Negative [Positive] Bias Temperature Instability, or NBTI [PBTI], and both are partially reversible on the removal of stress. Incorporating these recovery effects, the long-term degradation depends on the average duty cycle, but is independent of stressing signal frequency.

The overall effect of BTI is to reduce the maximum operating frequency, f_{Max}, of a circuit over its lifetime. To ensure that a chip meets its timing requirements over its lifetime, compensation techniques have been developed. In the presilicon design, appropriate delay guardbands may be added [1], [2], while the postsilicon phase may adapt the circuit during operation in the field [3], using sensors built in at the presilicon phase, by adjusting its clock frequency, supply voltage, or body bias. However, by definition, presilicon techniques are unaware of the runtime operating environment experienced by a chip and must consider worst-case scenarios by assuming pessimistic stress conditions for the circuits. Postsilicon techniques limit pessimism and deploy just enough adaptive compensation, based on monitors that periodically evaluate f_{Max} in the circuit under test (CUT). Two classes of monitors may be employed:

- *Surrogate circuit monitors*: These are test circuits used to estimate f_{Max} degradation in the CUT by trying to emulate the operating conditions/functionalities of the CUT. These range from simple ring oscillators (ROSCs) [4], [5] to more complex representative critical path (RCP) circuits [6], [7], [8].
- *CUT monitors*: In these methods, delay tests are directly performed on the CUT at predetermined intervals to measure its performance in terms of f_{Max} degradation [9], [10].

CUT monitors are accurate since they directly monitor the CUT, but may suffer from large hardware and test time overheads. Although such tests are required infrequently and their runtime can be reduced [11], the overheads of testing the entire chip may still be onerous. In this work we use surrogate circuit monitors (ROSCs to be specific) to characterize aging in the CUT.

Our contributions are summarized as follows. *First*, we propose a new *Upperbound on* f_{Max} (UofM) model to estimate a safe f_{Max} that the CUT can operate at. 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, by finding an envelope for the CUT delay. *Second*, we analyse the maximum pessimism in the UofM model and demonstrate it to be practically less than 0.2% in representative benchmark circuits. *Third*, we leverage the UofM model to present a novel aproach for inferring the delay degradation of the CUT based on data from on-chip ROSCs. Our scheme involves an initial presilicon characterization that uses a compact on-chip look-up table to determine a calibration factor, which we call the *degradation ratio*, $\mathcal{D}$, that translates ROSC measurement data to CUT delay degradation while capturing process-voltage-temperature (PVT) fluctuations in the manufactured circuit. Our solution accounts for BTI recovery when the circuit is power-gated and is robust to changes in the on-chip temperature and DVFS-related supply voltage changes. Our approach also captures the effects of process variations on the sensors and the CUT.

The rest of the paper is organized as follows. We begin with an explanation of on-chip monitors in Section II. Next, Section III presents a brief background on BTI-induced delay degradation, followed by detailed overview of the UofM model in Section IV. Section V shows the maximum pessimism in delay estimation possibly incurred by the UofM model using a given library of gates. Section VI demonstrates the experimental setup and results, and we conclude in Section VII.

II. Surrogate Circuits For On-Chip Monitoring

ROSC-based surrogate circuits are widely used in industry to evaluate aging. In silicon odometers [4], they have been demonstrated to provide high resolution and remove common-mode disturbances. ROSCs have several advantages over RCPs. *First*, ROSCs are compact and uniform, and require the design and layout of only a single repeatable macrocell, as against RCPs, which must be designed and laid out individually. *Second*, the process of generating RCP circuits is computationally expensive (for circuit b15, RCP generation can take 30 minutes [7], as against this work, where the computational effort for ROSC characterization takes less than two seconds).

However, like RCPs, ROSCs are mere surrogates for the CUT. Therefore, measuring f_{Max} degradation in ROSC is not equivalent to measuring degradation in the CUT, for several reasons. *First*, since the gate types on a CUT path are not the same as those on the ROSC, the path delay sensisitvity to V_{th}-shifts under aging is different from the ROSC delay sensitivity. Although RCP circuits try to overcome this issue, they suffer from the limitations of irregularity in layout and high design effort, as pointed out above, and are less used in industrial designs than ROSCs. *Second*, the ROSC has a single path that ages along a constant profile through its lifetime; in contrast, the delay of a CUT is the maximum of all path delays. Since a set of near-critical paths may have different aging sensitivities, the critical path may change over the lifetime of the CUT, causing it to age at different rates at different times.

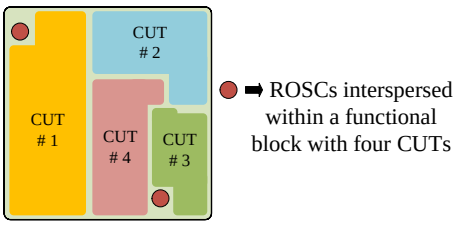


Fig. 1. Block diagram of a functional block with CUTs and ROSCs.

Within a larger circuit, ROSCs can be placed close to the CUT as illustrated in Fig. 1: since ROSCs are cheap and compact, many

copies can be replicated within the chip. Small circuit blocks may share a ROSC, while a very large block could contain several ROSCs.

To correlate well to the CUT, a test structure should try to match its (a) temperature, (b) V_{dd}, (c) process parameter variations, and (d) signal stress probability. Spatial proximity of the ROSC and the CUT ensures similar temperatures and enables the design to connect the supply of the ROSC to that of the CUT (thus capturing the effects of V_{dd} change under DVFS and power gating). Further, due to proximity, the ROSCs face a similar set of systematic process variations and spatially correlated random variations, i.e., shifts within the CUT in any manufactured part are similar to those in the nearby ROSCs. The granularity at which these sensors are deployed reflects a trade-off between overhead and accuracy: in this work, we assume that ROSC placement is a user input. Thus, it is easy to match criteria (a) and (b) and some parts of (c) above.

Matching random parameter variations and signal stress probabilities is harder since ROSCs are surrogate circuits. Shifts in delay due to purely random process variations in the CUTs are, by definition, physically impossible to capture in any surrogate; however, these effects can be diluted by using ROSCs with many stages [4]. Within CUTs, which typically have ten or more stages of logic on their critical paths, there is also a natural level of dilution of random variations. The switching activity of the CUT is similarly impossible to capture in a surrogate ROSC circuit. Therefore it is common practice to assume pessimistic worst-case stress probabilities for the CUT that will guarantee correctness of a prediction. Note that this pessimism is not specific to our method and is an unavoidable guardband that must be built into any design methodology based on surrogates [2]. In addition, we assume that sufficient margin has been kept for the clock skew due to aging in the clock network, since CUT aging is interpreted in terms of timing violation.

In our work, we use identical ROSC-based sensors, thus simplifying layout and design effort to insert them in the chip, and derive analytical methods to predict delay degradation (and hence operable f_{Max}) in the CUT based on measurements from the ROSC. Our method captures PVT effects due to temperature, V_{dd}, and systematic/spatially correlated variations correctly, dilutes the effects of random variations, and uses routine pessimistic approaches to address the built-in inability of ROSCs in capturing CUT stress probabilities.

III. The Impact Of BTI-induced Aging On Delay

The precise mechanisms of BTI are a matter of debate within the research community. Two candidates have emerged as the most important: the reaction-diffusion (RD) model [12] and the charge trapping (CT) model [13]. The key difference is in the trap generation mechanism in the oxide layer: the former [latter] gives a power-law [logarithmic] dependence of the V_{th}-shift with time.

In general, the V_{th}-shift depends both on the stress time and the average duty cycle. If the precise switching probabilities in a circuit are known, they may be used to determine the average duty cycle. However, in many cases, it is impossible to guarantee that specific switching probabilities will be maintained, and instead a worst-case probability may be used. Common approaches include using a worst-case constant-stress assumption on each transistor for NBTI, or a stress probability (SP) of 0.95 on each transistor [2].

The magnitude of the V_{th}-shift in PMOS and NMOS device at time t, $\Delta V_{th,p}(t)$ and $\Delta V_{th,n}(t)$ respectively, is given by:

$$\begin{aligned} \Delta V_{th,p}(t) &= K_1\ \xi_1\ f(t) = c_1 f(t) \\ \Delta V_{th,n}(t) &= K_2\ \xi_2\ f(t) = c_2 f(t) \end{aligned} \tag{1}$$

where ξ_1 and ξ_2 are the SPs of the PMOS and NMOS device, respectively, and K_1 and K_2 are constants dependent on temperature and V_{dd} specified by the aging model ([12] or [13]). Since PMOS [NMOS] devices are stressed when signal is low [high], ξ_1 [ξ_2] is the probability that the signal is at logic 0 [1]. The functions $f(t)$ are governed by the trap generation mechanisms for NBTI and PBTI. In principle, the functions $f(t)$ could be different for PMOS and NMOS devices, but these are experimentally observed to be the same, as documented in design manuals and the published literature [14]. Typically, $K_2 < K_1$.

It is particularly important to note that $f(t)$ is a *sublinear* and *monotonically increasing* function that captures BTI degradation; for the RD model, $f(t) \sim t^n$, where $n \sim 0.1 - 0.2$, and for the CT model, $f(t) \sim \log(t)$. Although the V_{th}-shifts through multiple stress-recovery cycles are not monotonic, $f(t)$ captures the envelope of the delay function, including recovery effects. The motonoticity property of the envelope is used later in Theorem 1.

For a given logic gate, let its delay $D(t)$ at time t be represented by the function $g(V_{th_x}(t))$ ($x : p$ or n depending on whether rise or fall delay is considered). For convenience, we henceforth drop subscripts p and n. Similarly, instead of c_1 and c_2, we use a general c and represent $\Delta V_{th} = c\ f(t)$. Under a small V_{th}-shift, the gate experiences a delay shift as:

$$\begin{aligned} D(t) &= g(V_{th}(t_0) + \Delta V_{th}) = g(V_{th}(t_0)) + \left.\frac{\partial g}{\partial V_{th}}\right|_{V_{th}(t_0)} \Delta V_{th} \\ &= D(t_0) + k\ f(t) \end{aligned} \tag{2}$$

where $D(t_0)$ is the delay of the gate at time t_0[1] and k is a constant multiplicative factor. Here, $k = Sc$, where S is the sensitivity of delay with respect to the absolute value of V_{th}, calculated at the nominal $V_{th}(t_0)$. Thus, under fixed stress conditions of temperature, supply voltage, and duty cycle, the delay is a function of time and is easy to compute as long as the nominal delay and sensitivity to V_{th} variation for each gate have been characterized.

IV. Delay Estimation And Aging Prediction

One of the primary difficulties in using a ROSC to predict the delay degradation of the CUT is that the CUT may have several near-critical paths that age at different rates and may become critical at various time points during its lifetime, while the frequency of the ROSC is determined by a single path. In this paper, we develop a method called the *Upperbound on f_{Max} (UofM)* procedure that provides a guaranteed upperbound on the delay (and thus the degraded maximum frequency of operation) of the CUT based on ROSC sensor data.

For an n-input gate, let D_i be the arrival time at input $i \in 1, \cdots, n$ of the gate, and $d_{i\to o}$ be the delay from input i to the output o of the gate; both parameters vary with time due to aging-related slowdowns. The arrival time at the gate output is given by:

$$D_o(t) = \max_{1 \le i \le n} \left(D_i(t) + d_{i\to o}(t)\right) \tag{3}$$

where the form of $d_{i\to o}(t)$ is given by Equation (2). Therefore, the arrival time at the output of the gate is the envelope of a set of delay curves corresponding to each argument of the $\max$ operator above.

If we perform static timing analysis (STA) over the entire CUT, we can obtain the temporal delay of the CUT: this is an envelope of a set of path delays (exemplified in Fig. 2 for a CUT with four paths). The idea of the UofM method is simple: if the max operator could be pessimistically approximated by a smooth function with continuous derivatives, such as the red curve in the Fig. 2, then the unitary operation of finding a smooth approximation to $D_o(t)$ at any gate output could be repeated to find a smooth approximation to the delay of the CUT.

[1]Normally, one might consider $t_0 = 0$, but realistically, chips undergo a burn-in phase causing some level of accelerated aging. Thus, based on the $f(t)$ functions characterized on fresh devices, we begin with a general value of t_0 as 3 months (as assumed in several prior papers).

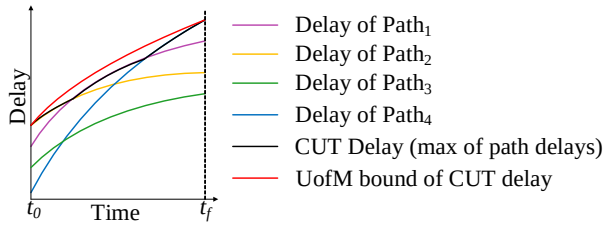


Fig. 2. Delay of CUT over its lifetime t_f.

A. An upper bound on the maximum delay

In this section, we lay the basis for the task of determining a smooth bound on $D_o(t)$ through the major theoretical result in our paper. Theorem 1 presents the case for upper-bounding the maximum of n aging curves. Pictorially, the theorem provides a precise expression for the red curve in Fig. 2, which is a continuous upper-bounding function for the maximum of n aging curves.

Theorem 1 In the interval $[t_0, t_f]$, an upperbound 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

$$y_n(t) = x_M(t_0) + \left[\frac{x_M(t_f) - x_M(t_0)}{f(t_f) - f(t_0)}\right](f(t) - f(t_0)) \quad (4)$$

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}.

Intuitively, the bound is simply the curve of the form in Equation (1) that matches the maximum curve at times t_0 and t_f. Note that the maximum at these two times could lie on different x_i curves, as illustrated in Fig. 2. The formal proof is deferred to Appendix I.

B. Applying the UofM bound to circuits using block-based analysis

In any circuit, the delay of a path is the sum of a set of gate delays whose temporal variations are each given by an equation of the form of Equation (2). Therefore, if $D_p(t)$ represents the delay of a path p in the circuit, the relationship between the path delays at any two times t_0 and t are given by a function of the form:

$$D_p(t) = D_p(t_0) + k(f(t) - f(t_0)) \quad (5)$$

From Equation (5), we can see that each path delay is similar to the form of the x_i functions in Theorem 1. Moreover, as discussed in Section III, the aging function $f(t)$ is a monotonically increasing function. Therefore, for a circuit with n paths, it is conceptually possible to obtain an upper bound, $y_n(t)$, on the delay of the circuit using the results of Theorem 1. As stated in Section III, while obtaining the UofM bound, we have considered SPs, $\xi_1 = 0.95$ and $\xi_2 = 0.95$, corresponding to NBTI and PBTI, respectively, at every gate of the CUT. Superficially, this may seem erroneous since a signal cannot be simultaneously high and low with the same probability. However, any input-output path through a gate goes through either the PMOS or the NMOS but *not* both. Therefore, either NBTI- or PBTI-based degradation is propagated, and our assignment of ξ_1 and ξ_2 correctly captures the worst-case delay for this worst-case path.

Theorem 1 allows the use of block-based STA by evaluating timing at the initial time t_0 and at the final time t_f, using the V_{th} aging model. In other words, only two timing analyses are needed to predict a safe upper bound on the delay over the entire lifetime.

The unitary operation in STA is to compute the maximum arrival time at the output of a gate, given the arrival times at the inputs. As an invariant, we assume that every arrival time is of the form in Equation (5): this invariant is preserved by writing the output arrival time in the same form. As is well known and apparent from Equation (3), STA involves two operations, "sum" and "max" and we now consider the preservation of the invariant under each operation.

- **Sum:** We add delay functions $D_1(t), \cdots, D_m(t)$, where $D_i(t) = D_i(t_0) + k_i(f(t) - f(t_0))$, to obtain

  $D_{sum}(t) = \sum_{i=1}^{m} D_i(t) = D_{sum}(t_0) + k_{sum}(f(t) - f(t_0))$

  where $D_{sum}(t_0) = \sum_{i=1}^{m} D_i(t_0)$ and $k_{sum} = \sum_{i=1}^{m} k_i$.
- **Max:** For the max over a set of delay functions $D_1(t), \cdots, D_m(t)$, each in the form of Equation (5), Theorem 1 immediately shows how the invariant is preserved.

Thus, block-based STA can compute the UofM function for the maximum arrival time at any node in the circuit, as well as the maximum delay of the circuit, in linear time in the number of gates.

C. Analyzing aging in the CUT based on ROSC data

Applying the analysis in Section IV-B to a given CUT, the temporal trend of the CUT delay can be represented using the UofM as:

$$D_{CUT}(t) = D_{CUT}(t_0) + k_{CUT}\,(f(t) - f(t_0)) \quad (6)$$

1) The delay of an aged ROSC: 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 and a fall delay of d_f, the period of the ring oscillator is well known to be $(2l + 1)(d_r + d_f)$. We refer to the period of a ROSC as its delay, D_{ROSC}, and express it as:

$$D_{ROSC}(t) = D_{ROSC}(t_0) + k_{ROSC}\,(f(t) - f(t_0)) \quad (7)$$

The change in period of a ROSC can be measured easily on-the fly by prestablished methods such as the silicon odometer [4], which uses the notion of *beat frequencies* to measure delay variations in the ROSC to a very high degree of precision.

2) The Degradation Ratio, $\mathcal{D}$: We now examine how ROSC aging measurements can be used to predict aging in the CUT. Let the delay degradation in the CUT at time t be given by $\Delta D_{CUT}(t) = D_{CUT}(t) - D_{CUT}(t_0)$, and let the corresponding value for the ROSC be $\Delta D_{ROSC}(t) = D_{ROSC}(t) - D_{ROSC}(t_0)$. From (6) and (7), we define the CUT *degradation ratio*, $\mathcal{D}$, as:

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

We make the following observations:

- From Equation (8), $\mathcal{D}$ for a CUT is a constant, independent of time t. Therefore, for any CUT, $\mathcal{D}$ can be precharacterized and stored in a look-up table to translate the ROSC delay degradation to the CUT delay degradation at various instants of time.
- The value of $\mathcal{D}$ may be different for different CUTs.
- If the CUT has one dominant critical path throughout its lifetime, the degradation ratio provides the true delay degradation of the CUT at any time (as does the UofM bound in this case).
- If the CUT has multiple critical paths that successively become dominant over its lifetime, $\mathcal{D}$ is based on the UofM bound, and Equation (8) provides a pessimistic bound on the CUT delay.
- PVT variations due to thermal and V_{dd} effects, systematic variations, and spatial correlations are all accounted for through the close spatial proximity of the ROSC and the CUT. Random variations in the ROSC can be reduced by using a larger number of stages [4].

D. Impact of temperature and V_{dd} on $\mathcal{D}$

As stated above, the effect of process variations is captured by the proximity of the ROSC and CUT, due to which their process parameters track each other. It is widely observed that a constant value of k in Equation (2) captures variations at all process corners. However, k depends on temperature, T and V_{dd} effects.

In this section, we investigate the impact of T and V_{dd} on $\mathcal{D}$ defined in Equation (8). To do so, it is necessary to analyse the $f(t)$

term in the aging model. The threshold voltage degradation with time for the RD model is given by [12] as:

$$\Delta V_{th}(t) = k_1 e^{\frac{E_{ox}}{E_0}} e^{\frac{-k_2}{T}} t^n \tag{9}$$

For the CT model, this is given by [15] as:

$$\Delta V_{th}(t) = k_3 e^{\frac{-k_4 V_{dd}}{T}} e^{\frac{k_5}{T}} [A + \log(1 + Ct)] \tag{10}$$

Here, $E_{ox} = \frac{V_{gs} - V_{th}}{t_{ox}}$, and $V_{gs} = \pm V_{dd}$ (during stressed/relaxed mode) and k_1, k_2, k_3, k_4, k_5, and E_0 are constants obtained from the aging model. Substituting $\Delta V_{th}(t)$ in Equation (2) and adding up gate delays to find the circuit delay degradation $\Delta D(t)$, we get:

$$\text{RD: } \Delta D(t) = k' e^{\left(\frac{-k_2}{T} + \frac{E_{ox}}{E_0}\right)} (t^n - t_0^n) \tag{11}$$

$$\text{CT: } \Delta D(t) = k'' e^{\left(\frac{k_5 - k_4 V_{dd}}{T}\right)} \log\left(\frac{1 + Ct}{1 + Ct_0}\right) \tag{12}$$

Here, k' and k'' represent the effect of adding the contributions of gate delays on a path during STA. The other terms are dependent on T, V_{dd} and time, which are identical for the CUT and ROSC by construction, and they cancel out when computing the ratio $\mathcal{D}$. Therefore, under both the RD and CT model, the value of $\mathcal{D}$ is:

$$\text{RD: } \mathcal{D} = \frac{k'_{CUT}}{k'_{ROSC}}; \qquad \text{CT: } \mathcal{D} = \frac{k''_{CUT}}{k''_{ROSC}} \tag{13}$$

Since the right hand sides of both equations above are independent of T and V_{dd}, the degradation ratio $\mathcal{D}$ is independent of T and V_{dd}.

V. Bounding The Maximum Pessimism In UofM Model

The UofM bound is a pessimistic estimate of the delay of a CUT even when the actual switching activity of the CUT is known. In this section, we present Theorem 2, which bounds the maximum pessimism incurred by the proposed UofM model. The proof of the theorem is provided in Appendix II.

Theorem 2 If k_1 and k_2 are the aging sensitivities of the gates in the current library with minimum and maximum percentage degradation over their lifetime, respectively, the maximum error in delay estimation in a CUT incurred by the UofM model using this library is upper-bounded by E_{max} as:

$$E_{max} = \frac{(k_2 - k_1)(f(t_f) - f(t_0))}{4} \tag{14}$$

where t_0, t_f, $f(t)$ and n have been defined earlier.

The maximum fractional error, E_{frac}, which is the ratio of E_{max} (Equation (14)) to $C_{top}(t')$ [or $C_{bot}(t')$] (delay of the two paths with maximum and minimum sensitivities when they cross over as shown in Fig. 5), is obtained by algebraic manipulation as:

$$E_{frac} = \frac{4(E_{max}^2)}{d_1 d_2 (\frac{k_2}{d_2} - \frac{k_1}{d_1})(f(t_f) - f(t_0))} \tag{15}$$

where d_1 and d_2 are defined in the proof of Theorem 2. The values of k_1, k_2 and d_2 can be found from the current library and d_1 obtained by adjusting the number of gates with minimum percentage delay degradation such that $d_1 - d_2 = 2E_{max}$. Thus, E_{frac} depends on the gate library, the numerical value of which is shown in the Section VI.

VI. Experimental Setup And Results

The ideas in this paper are exercised on a set of representative ISCAS'89 and ITC'99 benchmarks. The circuits are implemented using a gate library that consists of the following functionalities: two- and three-input NAND and NOR gates, three- and four-input AOI gates, inverter, and buffer, each with drive strength X1, X2 and X4, from the NanGate 45nm Open Cell Library. Each gate is characterized for nominal delay, output slew and delay sensitivities to V_{th} (for both rise and fall transitions) using the 45nm Predictive Technology Model (PTM). The benchmark circuits are synthesized using Synopsys Design Compiler.

The constant c_1 in Equation (1) is calibrated such that $V_{th,p}$ of the PMOS degrades by 25% in 10 years under a V_{dd} of 1.1V at an operating temperature of $85°$C. The function $f(t)$ in Equation (2) follows the power-law model (with $n = \frac{1}{6}$), and the constant c_2 is chosen so that the $V_{th,n}$ degradation in NMOS due to PBTI is one-third of that due to NBTI [3]. To account for the initial transient and burn-in, we set $t_0 = 3$ months and constrain the circuit lifetime to be 10 years beyond this point. The choice of burn-in period does not however affect our proposed methodology. In addition, since the actual SPs are unknown, we use a pessimistic value of 0.95 for both ξ_1 and ξ_2 (in Equation (1)) at every gate input of the CUT to obtain the corresponding k_{CUT} values, similar to [2] (however, our results will look fundamentally similar even if we use a worst-case SP of 1.0 instead of 0.95). We obtain k_{ROSC} using a 33-stage inverter chain as in [4], by considering SP of each inverter in the chain as 0.5. It is to be noted that we do not solve any placement problem of the ROSC in this work and assume them to be in enough proximity to the CUT so that perfect correlation exists between the process parameters of the ROSC and the CUT.

Table I presents the results of our method on the representative benchmarks at $T = 85°$C and $V_{dd} = 1.1$V. Each row corresponds to a single benchmark circuit associated with a single ROSC, except the circuit s_{mult}, which corresponds to a single ROSC that is shared by circuits s5378, s13207, and s15850. For each CUT, the second column in the table lists its gate count, $|G|$ and the third column shows the logical depth of the critical path d_{crit}: since the critical path may change over time, for convenience we consider the critical path at time $t = t_f$. We have observed that even if the critical path changes, the logical depth of the critical path does not vary appreciably over time. Generally speaking, larger values for d_{crit} correspond to larger values for the degradation ratio, $\mathcal{D}$, listed in the fourth column (note that $\mathcal{D}$ has no units).

Thus, we observe that extent of aging in a CUT is not necessarily dependent merely on the total number of gates, but also on the properties of the critical path(s), such as the logical depth. The sensitivity of a cell typically depends on its driving power and its load: cells with a larger driving power tend to have lower sensitivities, and those with larger loads have higher sensitivities. Critical paths are observed to contain large-sized cells, which have low sensitivity to V_{th}-shifts, but the cells that drive these large cells see large loads, particularly if they are smaller, making them sensitive to aging degradations. Thus, for critical paths with a larger number of stages, the impact of the larger cells is diluted by the smaller cells; this is less so for circuits with fewer stages. This explains why k_{CUT} (and hence $\mathcal{D}$, since k_{ROSC} is constant) generally increases with d_{crit}.

The next column represents the maximum percentage error of the UofM bound (over the entire lifetime): the estimation error due to the use of the upper bound is seen to be virtually negligible, even though each circuit has multiple near-critical paths. This is followed by a column that provides the percentage area overhead, ΔA, of the ROSC, i.e., the ratio of the ROSC area to the total area of the CUT+ROSC, as determined by Design Compiler. As expected, this area overhead is significant only for the smallest circuits, and is quite low when the CUT is larger. For the circuit, s_{mult}, in which the first three circuits in the table share a single ROSC, the overhead is small, and the estimation error is negligible. The final column lists the total CPU time, τ, that is required to compute the value of $\mathcal{D}$ for each CUT, evaluated on a 64-bit Ubuntu server (Intel® Core™2 Duo CPU E8400 3GHz). The modest runtimes (significantly faster than RCP procedures in, e.g., [7]) indicate the aptness of our method for

handling very large CUTs. Thus, during design time, these $\mathcal{D}$ values can be computed cheaply and stored in look-up tables for each CUT, using which one can estimate its delay degradation at any point of time based on a cheap measurement of ROSC delay degradation.

TABLE I
RESULTS FOR ROSC-BASED ESTIMATE FOR $T = 85^oC$ AND $V_{dd} = 1.1V$

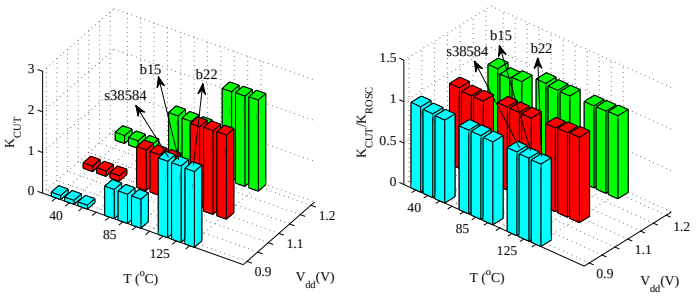
CUT	$\|G\|$ #	d_{crit} #	$\mathcal{D}$	Error (%)	ΔA (%)	τ (s)
s5378	690	11	1.51	0.17	4.63	0.50
s13207	590	18	1.88	0.00	5.74	0.46
s15850	336	18	1.95	0.00	10.41	0.36
s_{mult}	1616	18	1.95	0.00	2.06	0.78
s38417	4565	25	2.79	0.09	0.64	1.75
s38584	4585	26	2.61	0.00	0.67	1.69
b15	6311	63	6.35	0.00	0.49	1.99
b17	17882	62	6.28	0.00	0.17	5.18
b18	67776	130	12.78	0.00	0.04	16.95
b19	128494	120	11.89	0.00	0.02	31.64
b20	24080	128	12.25	0.00	0.13	6.09
b22	36149	128	12.23	0.00	0.09	8.38

To investigate the reason for the low errors, we further examined the characteristics of the critical paths. All of these circuits contain multiple near-critical paths, and although we do observe a crossover where the critical path changes as the circuit ages, the UofM bound is very close to the envelope of the maximum delay over time. We evaluate the bound in Theorem 2 to evaluate the theoretical maximum on the error. The maximum and minimum percentage delay degradation, $\left(\frac{D(t_f)-D(t_0)}{D(t_0)}\right)$, occur for the three-input NOR gate and buffer (having sensitivities k_2 and k_1), respectively, both with drive strength X4 and denoted as NOR3_X4 and BUF_X4. We synthesized a circuit with just two critical paths where each path was obtained by concatenating NOR3_X4 and BUF_X4 respectively. The nominal delays d_1 and d_2 of the paths were tuned by the number of stages in each path while keeping k_1 and k_2 unchanged, to obtain the maximum error scenario shown in Fig. 5. It can be proven that the maximum fractional error E_{frac} (as defined in Equation 15) for this circuit is the absolute maximum possible in this library (proof omitted due to space constraints). This error, found to be 3.59%, corresponds to the maximum possible pessimism of the UofM based delay estimate under our library.

In fact, achieving this bound requires a pathological case (since gates with small delays also tend to have small sensitivities) in which the critical path at t_0 has a low delay sensitivity, the critical path at t_f has a high delay sensitivity and is near-critical at t_0. The latter can be achieved if the critical path at t_0 has a small number of stages and the critical path at t_f has a large number of stages. This is unlikely to be seen in practice, and is not seen in any of our circuits. This is the reason why the UofM bound is even more accurate in practice than the already small bound on pessimism from Theorem 2.

Next, we evaluate the correctness of the notion that $\mathcal{D}$ is independent of the temperature, T, and V_{dd}, as claimed in Section IV-D. Simulations were run at various T (40°C, 85°C, 125°C) and V_{dd} (0.9V, 1.1V, 1.2V) values, and the dependence on the constants c_1 and c_2 in Equation (1) on V_{dd} and T was accounted for. Fig. 3(a) shows the k_{CUT} values of three CUTs, (s38585, b15, b22) for three values each of T on one axis and V_{dd} on another, normalized with respect to their baseline values at $T = 85°C$ and $V_{dd} = 1.1V$. For each (V_{dd}, T) point, the three bars correspond, from left to right, to s38584, b15, and b22, respectively. It can be seen that the bars at each such point are of equal height, indicating that each (V_{dd}, T) point experiences an equal multiplicative effect for each CUT.

Fig. 3(b) shows the $\mathcal{D}$ values, normalized to their corresponding baseline values in Table I, for the same three circuits and the T and V_{dd} values. When we examine the degradation ratio, $\mathcal{D}$, we find that all bars have a value that is very close to unity, i.e., $\mathcal{D}$ is independent



(a) Impact on k_{CUT} (b) Impact on $\mathcal{D}$

Fig. 3. Effect of change in temperature and V_{dd} on k_{CUT} and robustness of $\mathcal{D}$ to these changes

of V_{dd} or T. This may be understood by observing that k_{ROSC} also changes as T and V_{dd} are altered, and it tracks k_{CUT} very well at each point. The average absolute error of the normalized $\mathcal{D}$ from the ideal value of unity is only 3.79% for each of the three CUTs, across all T and V_{dd} values. Note that this error corresponds to a percentage of the already small delay shift (and not the delay), and is therefore negligible. This independence demonstrates that a single LUT serves the purpose of aiding delay degradation (and thus aging) estimation irrespective of variations in the operating temperature and V_{dd}, under our scheme where the CUT and the ROSC both operate in the same environment with regard to thermal changes and DVFS/power gating.

In this work, we are unable to show detailed comparisons with other approaches. To our knowledge, there is no other work that relates ROSC delays to CUT delays under aging. Existing work on RCPs for aging uses different methodologies, libraries and delay models. However, as pointed out earlier, the best proposed approach [7] requires significantly more computation than our method and entails complex layout issues. Moreover, our work is easier to implement in an industrial setting where ROSC-based methodologies have been in use for years.

VII. CONCLUSION

In this paper, we presented a technique to estimate delay degradation of a circuit using nearby ROSC-based aging sensors. We quantitatively determine how the data from the ROSC can be used to find the change in the circuit delay. Experimental results show that we can use the UofM metric to distill the relation between the CUT delay trend and the ROSC delay trend into a single degradation ratio, which can accurately predict the CUT delay degradation based on inexpensive measurements on the ROSC.

ACKNOWLEDGEMENT

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

REFERENCES

[1] S. V. Kumar, *et al.*, "NBTI-aware synthesis of digital circuits," in *Proc. DAC*, pp. 370–375, 2007.
[2] M. Agarwal, *et al.*, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. ITC*, pp. 1–10, 2008.
[3] S. V. Kumar, *et al.*, "Adaptive techniques for overcoming performance degradation due to aging in digital circuits," in *Proc. ASP-DAC*, pp. 284–289, 2009.
[4] 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.
[5] T. B. Chan, *et al.*, "DDRO: A novel performance monitoring methodology based on design-dependent ring oscillators," in *Proc. ISQED*, pp. 633–640, 2012.
[6] Q. Liu and S. S. Sapatnekar, "Synthesizing a representative critical path for post-silicon delay prediction," in *Proc. ISPD*, pp. 183–190, 2009.

[7] S. Wang, *et al.*, "Representative critical reliability paths for low-cost and accurate on-chip aging evaluation," in *Proc. ICCAD*, pp. 736–741, 2012.
[8] X. Wang, *et al.*, "Path-RO: a novel on-chip critical path delay measurement under process variations," in *Proc. ICCAD*, pp. 640–646, 2008.
[9] M. Agarwal, *et al.*, "Circuit failure prediction and its application to transistor aging," in *IEEE VLSI Test Symp.*, pp. 277–286, 2007.
[10] Y. Li, *et al.*, "CASP: concurrent autonomous chip self-test using stored test patterns," in *Proc. DATE*, pp. 885–890, 2008.
[11] Y. Li, *et al.*, "Concurrent autonomous self-test for uncore components in system-on-chips," in *IEEE VLSI Test Symp.*, pp. 232–237, 2010.
[12] S. Chakravarthi, *et al.*, "A comprehensive framework for predictive modeling of negative bias temperature instability," in *Proc. IRPS*, pp. 273–282, 2004.
[13] 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.
[14] 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.
[15] J. B. Velamala, *et al.*, "Physics matters: statistical aging prediction under trapping/detrapping," in *Proc. DAC*, pp. 139–144, 2012.

APPENDIX I

In this section, we present a proof of Theorem 1. We begin by presenting a lemma for a simpler version of the theorem for just two paths, and then prove the theorem.

Lemma 1 Consider two monotonically increasing functions $x_1(t)$ and $x_2(t)$ in the interval $[t_0, t_f]$.

$$x_1(t) = x_1(t_0) + k_1(f(t) - f(t_0))$$
$$x_2(t) = x_2(t_0) + k_2(f(t) - f(t_0)) \quad (16)$$

An upper bound on maximum of $x_1(t)$ and $x_2(t)$ is given by:

$$y_1(t) = x_M(t_0) + \left[\frac{x_M(t_f) - x_M(t_0)}{f(t_f) - f(t_0)}\right](f(t) - f(t_0)) \quad (17)$$

where the function $x_M(t) = \max_{i\in\{1,2\}}(x_i(t))$ represents the upper envelope of the functions x_1 and x_2.

Proof: Without loss of generality, assume that $x_1(t_0) \leq x_2(t_0)$. Note that since both curves are montonically increasing, one of two possibilities must be satisfied, as illustrated in Fig. 4.
Case I: If $x_1(t_f) \leq x_2(t_f)$, then x_2 dominates x_1 over the interval.
Case II: If $x_1(t_f) \geq x_2(t_f)$, the curves cross over once in $[t_0, t_f]$.

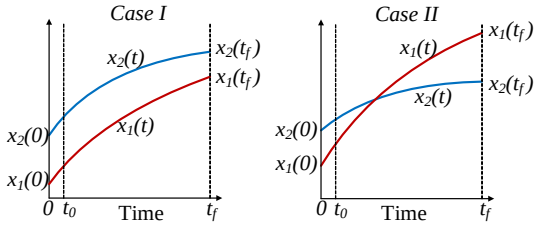


Fig. 4. Possible trends for monotonically increasing $x_1(t)$ and $x_2(t)$.

In Case I, the result is trivially true, since the expression evaluates to the equation for $x_2(t)$. For Case II, $x_M(t_0) = x_2(t_0)$, $x_M(t_f) = x_1(t_f)$. Observing that $k_1 = \frac{x_1(t_f) - x_2(t_0)}{f(t_f) - f(t_0)}$, and performing some algebraic manipulation, we find that:

$$y_1(t) - x_1(t) = (x_2(t_0) - x_1(t_0))\left[\frac{f(t_f) - f(t)}{f(t_f) - f(t_0)}\right] > 0, \quad (18)$$

$$y_1(t) - x_2(t) = (x_1(t_f) - x_2(t_f))\left[\frac{f(t) - f(t_0)}{f(t_f) - f(t_0)}\right] > 0, \quad (19)$$

Each line above evaluates to be positive due to the monotonicity of f, i.e., $f(t_f) > f(t) > f(t_0)\ \forall\ t_0 < t < t_f$. Thus, $y_1(t)$ is an upper bound on $\max(x_1(t), x_2(t))$ over the interval $[t_0, t_f]$. □

Proof of Theorem 1: Building upon Lemma 1, the proof is presented by mathematical induction over n. For the basis case, $n = 1$, Lemma 1 demonstrates that $y_1(t)$ forms an upper bound on $\max(x_1(t), x_2(t))\ \forall t \in [t_0, t_f]$, and has the form:

$$y_1(t) = x_M(t_0) + \alpha_1(f(t) - f(t_0))$$

where α_1 is a constant of the form as in Equation (4).

For the inductive step, we assume that $y_{n-1}(t)$ is an upper bound on maximum of $x_1(t), \cdots, x_n(t)$, and attempt to show that $y_n(t)$ is an upper bound on maximum of $x_1(t), \cdots, x_{n+1}(t)$.

From the inductive hypothesis, $y_{n-1}(t)$ is an upper bound on the first n functions with the form:

$$y_{n-1}(t) = x_M(t_0) + \alpha_{n-1}(f(t) - f(t_0))$$

where α_{n-1} is a constant of the form in Equation (4). Therefore it is enough to prove that $y_n(t)$ is an upper bound on the maximum of $y_{n-1}(t)$ and $x_{n+1}(t)$ for $t \in [t_0, t_f]$. This result follows immediately from Lemma 1. In particular,

$$x_M(t_0) = \max(y_{n-1}(t_0), x_{n+1}(t_0)) = \max_{1\leq i\leq n+1} x_i(t_0)$$

$$x_M(t_f) = \max(y_{n-1}(t_f), x_{n+1}(t_f)) = \max_{1\leq i\leq n+1} x_i(t_f) \quad \square$$

APPENDIX II

Proof of Theorem 2: Consider a CUT with multiple critical paths over its lifetime. Let us represent the paths which are critical at $t = t_0$ and $t = t_f$ by P_1 and P_2 and denote them by the curves $C_{bot}(t)$ and $C_{top}(t)$, respectively, (Fig. 5 exemplifies this for a CUT with three critical paths) as:

$$C_{bot}(t) = d_1 + k_1(f(t) - f(t_0))\ ;\ C_{top}(t) = d_2 + k_2(f(t) - f(t_0))$$

where d_1 and d_2 are the delays of P_1 and P_2 at $t = t_0$ respectively.

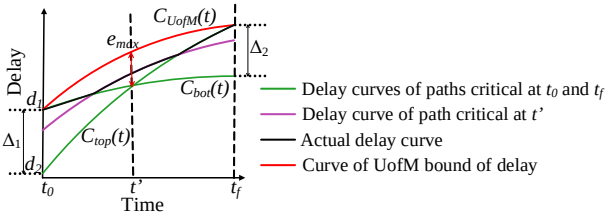


Fig. 5. Error bound for the UofM model.

Evidently, $d_2 < d_1$. Using the UofM model, the estimated delay, denoted by $C_{UofM}(t)$, (e.g., the red curve in Fig. 5) is given by:

$$C_{UofM}(t) = d_1 + \left(k_2 + \frac{d_2 - d_1}{f(t_f) - f(t_0)}\right)(f(t) - f(t_0)) \quad (20)$$

The deviations of $C_{UofM}(t)$ from $C_{bot}(t)$ and $C_{top}(t)$ are given by $e_1(t)$ and $e_2(t)$, respectively, as:

$$e_1(t) = \frac{f(t) - f(t_0)}{f(t_f) - f(t_0)}\Delta_2\ ;\ e_2(t) = \frac{f(t_f) - f(t)}{f(t_f) - f(t_0)}\Delta_1 \quad (21)$$

where Δ_1 and Δ_2 are the differences in the the two path delays at t_0 and t_f, respectively. The error (pessimism) of the UofM curve is bounded by minimum of $e_1(t)$ and $e_2(t)$ (as can be visualized from Fig. 5), which are monotonically increasing and decreasing, respectively. Therefore, the maximum error, e_{max}, occurs when both are equal, i.e., at $t = t'$ when the two curves cross over:

$$e_{max} = \Delta_1\left(1 - \frac{\Delta_1}{(k_2 - k_1)(f(t_f) - f(t_0))}\right) \quad (22)$$

Given the value of k_1 and k_2 as the sensitivities of the paths with minimum and maximum percentage degradation possible using the current gate library, the choice of Δ_1 can be optimized to Δ_1^{opt} that gives the maximum value of e_{max}, denoted by E_{max}. In other words, given fixed k_2 and d_2 for the path P_2, and k_1 for P_1, the number of gates in P_1 can be adjusted to obtain d_1 such that $d_1 - d_2 = \Delta_1^{opt}$. Note that changing number of cells in P_1 should not change k_1 (which can be ensured by concatenating same type of gates in the path). We obtain Δ_1^{opt} by differentiating e_{max} with respect to Δ_1, obtaining the maximum value of e_{max}, E_{max} as $(\Delta_1^{opt}/2)$, when $\Delta_1^{opt} = \frac{(k_2 - k_1)(f(t_f) - f(t_0))}{2}$. The result follows immediately.

At this optimum value, $\Delta_1 = \Delta_2$, so that the differences in the two path delays at time t_0 is identical (and the negative of) the difference at time t_f. □