Small Embeddable NBTI Sensors (SENS) for Tracking On-Chip Performance Decay

Adam C. Cabe, Zhenyu Qi, Stuart N. Wooters, Travis N. Blalock, Mircea R. Stan
University of Virginia, 301 McCormick Rd., Charlottesville, VA 22904, USA
{acc9x, jerry, snw7f, tnb2a, mircea}@virginia.edu

Abstract — On-chip circuit aging sources, like negative bias temperature instability (NBTI), hot-carrier injection (HCI), electromigration, and oxide breakdown, are reducing expected chip lifetimes. Being able to track the actual aging process is one way to avoid unnecessarily large design margins. This work proposes a sensing scheme that uses sets of reliability sensors capable of accurately tracking NBTI PMOS current degradations across process, temperature, and varying activity factors. We show that a set of 1000 such small sensors can predict chip lifetime to an uncertainty of 7% to 10%. We also show that, once the total area dedicated to sensing is chosen, the lifetime prediction uncertainty is almost insensitive to the tradeoff between the number of sensors and the area of each individual sensor.

Keywords — NBTI, Circuit Reliability, DFM

I. Introduction

Circuit designers are challenged with overcoming several aging effects when designing robust integrated circuits (IC) in modern ultra-scaled technologies. Circuit lifetime has become limited by physical reliability phenomena such as NBTI, HCI, electromigration, and oxide breakdown. These aging effects manifest themselves as changes in key circuit and device parameters. Furthermore, these degradations worsen with technology scaling, particularly as oxide thickness decreases, and current density, oxide field strength, and die temperature increase.

Recent works depict PMOS NBTI as the primary parametric failure mechanism in modern ICs [1]. NBTI stems from the generation of interface traps at the Si/SiO₂ boundary when stressed by a negative PMOS gate-source bias, and is exacerbated at elevated temperatures. The degradation manifests itself as a decrease in transconductance g_m and increase in the absolute value of PMOS V_th over time, eventually leading to critical performance degradation upwards of 20% [2].

One potential solution is to develop real-time system-level reliability enhancement techniques that can compensate for the degrading effects of NBTI and other similar phenomena [3]. These methods help prolong circuit lifetime by dynamically adjusting circuit and system parameters according to real-time feedback from on-chip sensors. Our work stresses the importance of such sensors, as they can provide both data to help understand the physics behind such aging effects, and feedback information for these real-time reliability enhancement techniques.

Several key metrics embody the efficacy of a sensor, including accuracy, precision, power, size, speed, and the ability to track with variations in temperature and activity factor. Designing an optimal sensor scheme involves numerous tradeoffs between these metrics. Our work contributes a novel reliability sensor, capable of accurately tracking NBTI degradations across mismatch, temperature, and activity factor variations. The sensor size is adjustable, making it embeddable near critical paths and blocks, thus allowing it to track the actual temperature and activity factor experienced by neighboring components.

The statistical nature of these sensors is discussed in detail, particularly considering how process variations impact the required size and number of sensors. Simulations show that given a total effective sensor coverage area, the tradeoff between the number of sensors and the area of one sensor has minimal impact on accuracy. Consideration is also given to how temperature impacts sensing accuracy, and affects the on-chip placement of such sensors.

Section 2 begins with an analysis of prior NBTI sensors. Section 3 follows by detailing the sensor circuits and timing, including device sizing impacts. Next, the sensor NBTI models, statistics and accuracy are discussed in sections 4 and 5. Section 6 gives insight into the sensor physical implementation and usage, providing a detailed circuit layout. Section 7 concludes this work.

II. NBTI Sensing Review and Goals

NBTI impacts circuits in several visible ways including decreasing PMOS drain-source current, and reducing circuit speed and leakage. Several groups have developed sensors to detect these changes [4], [5], [6], [7]. Some solutions employ ring-oscillator based sensors to track frequency degradations, attaining good accuracy while sacrificing area and complexity [6], [7]. The sensor in [6] is comprised of two 105-stage ring oscillators, and [7] deploys a 15-stage NAND ring oscillator biased in subthreshold mode.
Other monitors have been proposed to track non-frequency NBTI impacts like leakage reduction using $I_{DDQ}$ testing [5], and dynamic-locked-loop control voltage shifts [4]. These techniques are highly accurate, however they would be difficult to use in real-time monitoring systems since their tracking requires off-chip analog measurements.

Common to each of these sensing techniques is that they track aging constantly throughout the chip’s life-span. Fig. 1 shows a typical bathtub curve, plotting failure rate versus chip age. The aforementioned sensors can provide NBTI degradation data for any point on that curve, which can require a high degree of sensor characterization, calibration, and modeling.

Instead of providing data at every point on this bathtub curve, the goal of our work is to identify one or more critical points along the way. In essence, our sensors serve as a chip “check engine light,” informing the system when the circuits have degraded by a certain critical percentage $\Delta I$. Tracking to only certain critical points reduces the amount of information needed from the sensor, allowing smaller sensor sizes while still achieving good accuracy. As in other sensing schemes, multiple sensors are required to track over process variations, but the individual sensors can be smaller, thus allowing for increased system embeddability, and decreased total sensing area.

Fig. 1 shows some possible critical $\Delta I$ values, where sensors can be used to trigger and alert the system about various drain current changes. In this example, nine and ten percent are in the “end-of-life” region; however this is only a qualitative representation, and will change depending on the application. The sensors used to track these critical $\Delta I$ values are detailed in the next sections.

### III. Sensor Circuit, Timing, and Sizing

#### A. Basic Sensor Functionality

A simple block diagram of the proposed NBTI sensor is shown in Fig. 2, being comprised primarily of a degradation tracking inverter and a reference inverter. The sensor is simple, requiring no analog measurement circuitry or biasing voltages, and is more compact than previously demonstrated ones.

The degradation tracking inverter is designed to deliver initially a current slightly larger than the reference inverter, where the percentage margin is equivalent to $\Delta I$ as discussed in Fig. 1. This margin is simply set by upsizing the degradation tracking inverter’s PMOS device.

The sensor functions by nominally powering down the reference inverter so that it does not degrade, while running the tracking inverter at a specified switching activity $\alpha$. This switching input can be a replica of a critical path input so the sensor accurately tracks with that critical component, or it can be a set pulse with a known activity factor. Periodically, the inverters are connected in a cross-coupled configuration, and the internal node outputs (initially pre-discharged) are determined by the strength of the inverter PMOS transistors. Before any aging, the tracking inverter PMOS is stronger and will always pull its output high, however, once the inverter degrades beyond the set $\Delta I$, the reference inverter, which has not aged, is the stronger one and “triggers” the sensor output.

#### B. Circuit and Timing Analysis

Fig. 3 shows a transistor-level schematic of the NBTI sensor, containing two inverters, switches, and control signals. The internal nodes are read out through inverters, and pass transistors are used to control the connections between the reference and degrading transistors. The sensor has two operating modes, tracking and polling. The timing diagram in Fig. 4 shows the control signals and inverter outputs during these two modes, and is obtained from SPICE-level circuit simulations in a 65nm technology.

During tracking mode, the $ctrl$ signal is held high such that the degradation tracking inverter receives the desired input for tracking NBTI. As stated previously, this input can be a replica of a critical path input, or it can be a set pulse with a known activity factor. A primary benefit of our sensor over ring oscillator based designs is that it can track more than just a fifty percent switching activity; this increases the accuracy of the sensing solution. During tracking mode, the reference inverter is disconnected from the tracking inverter by the open pass gates, and its input is held at $V_{dd}$ to keep the PMOS device from degrading.

The polling mode is shown just to the right of the tracking mode in Fig. 4. During the first moments of polling, the $dischry$ signal is asserted, pre-discharging the input nodes to both inverters and ensuring that only the PMOS devices affect the inverter output levels. The inverters are
then connected in a cross-coupled configuration, and the stronger one pulls its output node high, assuming that the loads seen from the output of each inverter are equal. The tracking inverter is stronger in this example simulation, and its output latches to logic high. However over time, this inverter’s PMOS device will degrade and eventually the reference inverter will pull its output high, triggering the sensor and signalling that the critical $\Delta I$ has been reached.

### C. Impact of Transistor Sizing and Load Balancing

The sensor functionality and accuracy are critically impacted by the sizing of each transistor in the circuit. First, adjusting the ratio between $P_D$ and $P_R$ in Fig. 3 sets the circuits initial $\Delta I$ margin. Designing $P_D$ $x$ percent larger than $P_R$ should create close to a $x$ percent current ratio between the tracking and reference inverters, however second order effects will make some manual width adjustments required to achieve the exact desired margin.

A second impact of sizing is on sensor power, as is shown in Fig. 5. This plot details the power during both the polling and the tracking modes given a 1GHz tracking input as shown in Fig. 4. Power increases as plotted against the reference PMOS width, and this result indicates that more power is consumed during the polling mode, largely due to the two competing inverters both trying to pull their respective outputs high. The circuit generally operates in the tracking mode, placing it in the low power regime for a majority of its operating life.

Circuit sizing also determines how fair the fight is between the reference and degrading inverters. To make sure that the sensor is sensitive only to differences between the PMOS tracking and reference devices, it is imperative that each inverter’s output load be matched. To do this, the sensor is designed in a symmetric fashion, including equally sized tri-state buffers that provide inputs, equally sized NMOS pass transistor switches, and equally sized inverters to buffer the outputs.

Lastly, the transistor sizing impacts the current variations of each sensor, and consequently the accuracy of each sensor. It is well known that transistor variations scale by approximately $\frac{1}{\sqrt{W \cdot L}}$ [8], so increasing the area of a PMOS device decreases its percentage $I_D$ variation. The next sections discuss in detail how the device statistical distributions determine the overall accuracy of the sensing scheme, including the tradeoffs between total sensor area, size, and number of sensors.

### IV. Statistical Modeling and Analysis

#### A. Modeling Sensor Distributions with NBTI

The proposed NBTI sensor is comprised of two inverters, each having its own distribution of current based on its size. Fig. 6 shows distribution samples for a large (dashed curve) and a small (solid curve) sensor, where the current is normalized and the degrading inverter delivers 10% more current than the reference. The distributions show that larger sensors have less current variation, which is equivalent to saying that more sensing area, or more sensors, yield less variation. Random intra-die variations are considered the primary variation source, since the close spatial locality of the sensor inverters heavily reduces the impact of systematic intra-die variations.

The NBTI current decay is modeled in simulation by shifting the tracking inverter’s distribution over time. This rate is logarithmic with time, and is controlled by temperature, gate voltage, oxide thickness, and activity factor. The decay is modeled as a change in $V_{th}$, where the resultant $\Delta V_{th}$ stems from the diffusion and back-diffusion of $H$ and $H_2$ across the gate oxide [9].
NBTI is unique in that it experiences both stress and recovery phases, where turning the PMOS device off actually recovers some of the degraded $V_{th}$. This recovery depends on the activity factor and duty cycle of the PMOS input, and reduces in magnitude with scaling technology nodes [6], [1]. The $\Delta V_{th}$ formulas describing the stress and recovery periods of NBTI degradation are shown in Eqns. 1, 2, and 3, which are taken from [9].

$$\Delta V_{th} = \begin{cases} (K_v(t - t_0)^{0.5} + 2\sqrt{\Delta V_{th0}})^{2n} & \text{(stress)} \\ (\Delta V_{th0} \cdot (1 - \frac{1 - 2\xi t}{2\xi + \sqrt{\xi C(t - t_0)}})) & \text{(recovery)} \end{cases}$$

(1)

$$K_v = (\frac{q}{e})^3 K_f C_o (V_{gs} - V_{th}) \sqrt{C_e} \exp(\frac{2E_{ox}}{E_{ox1}})$$

(2)

$$C = \exp(-E_a/kT)/T_0$$

(3)

The fit parameters in Eqns. 1–3 are given and discussed in the original model work [9]. Results from this model shows $V_{th}$ to decay logarithmically with time, and to vary exponentially with temperature as shown in Fig. 7.

### B. Reaching the Critical $\Delta I$

The reference and degrading distribution variances change with the sensor area and the total number of sensors, and these distributions determine the sensing scheme’s accuracy. In general, larger area provides increased accuracy. Multiple sensors are typically used to track NBTI, mainly to keep each individual sensor’s area down and make it easier to embed them near dense digital components. Given multiple sensors, critical questions arise: how many sensors need to “trigger” to determine when the critical $\Delta I$ is reached, and how many sensors are needed for good accuracy?

For our design, a sensor “triggers” when the tracking inverter falls weaker than the reference. Polling each sensor provides a total count of $T$ triggers, where $T$ is binomially distributed with a probability of triggering $p = \frac{X}{N}$, where $N$ is the total number of sensors. The degrading and reference inverter threshold voltage distributions are both gaussian and independent, since only mismatch variations are considered due to the inverters’ close proximity to each other.

Polling the sensors is equivalent to selecting a sample from each distribution and comparing to see which has the larger current. The time when the margin $\Delta I$ is erased occurs when there is a one half probability that the reference distribution sample ($p_r$) is larger than the degrading distribution sample ($p_d$), or $p_d > p_r = 50\%$.

Intuitively, this can be described by picturing the two gaussian distributions as having the same mean when the margin is erased. Since gaussian distributions are symmetric, it follows that $p_d > p_r = p = 50\%$ when the two distributions means are equal. Furthermore, the binomial maximum likelihood estimate (MLE) says to maximize the probability $p$, the MLE of $p$ is $\hat{p} = \frac{X}{N}$ [10]. So to maximize the probability of $p = 50\%$ requires $\frac{X}{N} = \frac{1}{2}$, or half of the sensors to trigger. Therefore, when half of the sensors trigger, the likelihood that the critical $\Delta I$ is reached is maximized. It is interesting to note that this result is independent of the variance as long as the distributions are gaussian!

The next section continues to explore the relationship between the area of one sensor and the total number of sensors, analyzing their tradeoffs and their impact on accuracy.

### C. Tradeoffs Between Area and Number of Sensors

Intuitively, the same accuracy should be obtainable with fewer large sensors, or with more small sensors. This section attempts to quantify this intuition. Fig. 8 plots the number of sensors vs. accuracy. The accuracy is obtained by performing 1000 Monte Carlo simulations, and taking the difference between the longest and shortest times to trigger half of the sensors. Each curve assumes a fixed sensor size, and is based on transistors with a 1.2nm gate oxide, and a 1.2V $V_{gs}$ at a temperature of 370K. The PMOS device sizes range from .1 to 10µm, use three times minimum gate length of .18µm, and are always on (degrading). The distribution standard deviations are calculated from SPICE monte carlo analyses using an 65nm process design kit.
The displayed accuracies range from 1/one day for many large sensors, down to 1/one year for few small sensors. The critical percent change is set to 10%, placing the lifetime at approximately 100 days with constant degradation (with no recovery, this is only for illustration purposes, in real life recovery and non-100% activity would lead to longer lifetimes). This means that the accuracies achieved are between 1% and 350% of the total lifetime depending on the total sensing area.

The work in [11] states that a minimum of 1000 of their sensors are necessary to attain a 10% accuracy considering both oxide breakdown and NBni degradations. We do not know the exact size of the transistors in their sensors, but the result is qualitatively consistent with our simulations, where using 1000 13.96\(\mu\)m\(^2\) sensors yields a 9.5% accuracy, while employing 1000 14.2\(\mu\)m\(^2\) sensors achieves a 7% accuracy.

Fig. 9 goes further to examine the accuracy tradeoff between the area of one sensor and the total number of sensors. This plot shows the number of triggered sensors against time, considering four different total number of sensors while fixing the total sensing area. The two curves for each number of sensors represent the extremes within the monte carlo distribution, and the accuracy is taken as the difference in the two extremes when fifty percent of the sensors trigger. This accuracy is projected down onto the x-axis.

A key result from this analysis is that given a set total sensing area, the tradeoff between the number of sensors and the area of one individual sensor is essentially even, barring some slight margin for area overhead per sensor. This means that the total sensing area is primarily responsible for the accuracy of the sensing scheme, independent of how many sensors comprise this given area. For example, 1000 sensors like the ones described here provide approximately the same accuracy as 100 sensors where each sensor would be 10 times larger. The impact of total sensing area on accuracy is plotted in Fig. 10.

The total area shown here is comprised of the reference and degrading inverters, the tri-state buffers, and the output buffers for each sensor. The area of the output and tri-state buffers are taken from standard-cell layouts in an industrial 65nm process design kit. This plot shows that an accuracy of 10% requires a total sensing area of approximately 7000\(\mu\)m\(^2\), assuming a temperature of 370K. This is on par with other results, being about 3-4 times smaller than the sensing area discussed in [7], however the sensors in that work track to both dielectric breakdown and NBni.

The simulation results presented through the last several sections mostly assume a constant temperature through degradation. In reality, the temperature may intensely fluctuate across the die, causing different sensors to age at different rates. The next sections discuss temperature, looking at both how temperature impacts the accuracy, and discussing sensor placement strategies faced with temperature hot-spots.

V. Considering Temperature Variations

The model equations presented in Sec. 4A showed NBni to be exponentially dependent on temperature. Fig. 11 plots the accuracy, relative to lifetime, against temperature, where the sensing area ranges from 1000 to 30000\(\mu\)m\(^2\). This simulation shows that although temperature changes the total lifetime of a sensor, it doesn’t affect the relative predictive accuracy to lifetime.

Temperature fluctuations can reduce the accuracy of any NBni sensing scheme, however measures can be taken to limit these negative impacts. An architectural approach to this problem might be to first analyze the thermal map of
the chip using a tool such as HotSpot [12]. With a chip-level temperature map, it is possible to place sensors in areas of relatively uniform temperature, reducing the overall temperature gradient seen by the sensors, and therefore lessening the impact on accuracy.

Furthermore, temperature directly impacts the choice of critical $\Delta I_d$ designed into the sensors. NBTE decays exponentially with temperature, and for low temperatures particularly near room temperature, PMOS devices may only experience current decreases of 5-7% during their lifespan. The critical $\Delta I_d$ designed into the sensor should be tailored to the expected operational temperatures.

VI. Layout Design

Until now we established the even tradeoff between the area of one sensor and the number of sensors, given a fixed total sensing area. A benefit of this is that it allows the sensor size to be custom tailored based on its on-chip location. This means that the sensor size should be easily adjustable in order to meet different spatial and margin $\Delta I_d$ requirements. Furthermore, major consideration is placed on reducing the variation between each inverter in attempts to maximize the sensing accuracy.

Fig. 12 shows the sensor layout designed in a 65nm technology. This layout shows only the inverter sensing areas, as the other blocks are commercially licensed standard cells. The sensor height of 3.2µm is constrained by the height of surrounding standard cells, which allows this sensor to easily abut into an automatic placement ASIC design flow. The sensor overhead from the tri-state buffer and inverter standard cells adds 13.6µm² per sensor, and is constant over active sensing area size.

The degradation tracking and reference PMOS devices are laid out with fingers to allow for easy area changes. However, since these devices are large, there is a high chance of lateral variation between fingers. This variation is minimized by using interdigitation, which involves alternating fingers of two transistors in layout [13]. An additional step to reduce variation is to increase the gate lengths to three times the minimum feature size.

VII. Summary and Future Work

This work introduced a novel NBTE sensor, designed to track current degradations with temperature and varying activity factors. The sensing scheme achieved accuracies of better than 10% given a total sensing area of at least 7000µm². Sensor layouts showed steps taken in physical design to reduce the variation impacts.

As future work we plan to implement these sensors at the system level. This will require handling the control signals for the two modes of operations, and counting the number of triggers when polling. A finite state machine could handle the required three control signals. Several methods could be used to tally the number of triggered sensors, including a scan-chain or a parallel majority voting scheme. Since speed is not of primary concern, a simple scan-chain should accomplish the task, where the tallies are added together with a simple counter circuit.

Acknowledgment

We thank Chris Gregg and Wei Huang from the University of Virginia for illuminating discussions on this topic. This work was supported in part by a grant from Intel, by an NSF CRI (CNS-0551630) grant, and by the Interconnect Focus Center, one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation and DARPA program.

References