Journal of Electronic Testing (2019) 35:87–100
https://doi.org/10.1007/s10836-019-05772-5

An Efficient Metric-Guided Gate Sizing Methodology for Guardband Reduction Under Process Variations and Aging Effects

Andres Gomez[1,2] · **Victor Champac**[1]

Received: 1 September 2018 / Accepted: 4 January 2019 / Published online: 22 January 2019

Abstract

Circuit reliability due to Bias Temperature Instability, BTI, has become an important concern in scaled-down complex electronic systems. Even more, current silicon technologies are severely affected by the combined impact of BTI-induced device's aging and Process-induced device's parameters variations. The conventional worst-case guardbanding to deal with reliable circuit operation is not longer an efficient approach as the circuit performance is significantly penalized. This paper presents a gate-sizing optimization me-thodology to reduce the worst-case guardbanding considering the combined effects of aging due to BTI and process variations. The proposed methodology allows to trade-off the reduction of guardbanding and the area cost. The proposed methodology uses multiple workload-aware aging analysis procedures to identify a realistic workload condition that causes maximum degradation to each potential critical paths of the circuit. In such a way, classic worst-BTI assumptions that lead to over-design are avoided. New gate-sizing metrics are proposed to identify the most beneficial gates to resize in the delay optimization process. In order to compute the gate sizing metrics efficiently, it is proposed a fast approximation for the sensitivity of the statistical delay of a path with respect to a change in the size of a gate. Also, the criticality, slack-time and area penalization are considered in the metric. A heuristic is proposed to guide the iterative delay optimization process. Some key conditions are identified in the workload analysis, metric evaluation and the heuristic to reduce the computational cost. The results show clearly the benefits of using multiple workload-aware aging analysis and the proposed gate-sizing metrics. It is shown that the proposed gate-sizing metrics are more efficient than others available in the literature since they provide a better area-guardband reduction trade-off. The proposed methodology results in more reliable designs at low area overhead, and it is suitable to guarantee the stringent quality requirements of modern circuits.

Keywords Aging of circuits and systems · Statistical timing analysis · Design optimization · Gate sizing metrics

1 Introduction

As technology scales down device's feature size, circuit's lifetime reliability has become a major challenge in integrated circuit design, mainly due to transistor aging induced by Bias Temperature Instability (BTI) mechanism [1]. BTI causes a gradual increase on the device's threshold voltage (V_{th}) over the lifetime, increasing delay, and ultimately, it can make a circuit to violate time specifications. The impact of BTI on circuit delay degradation (and lifetime reliability) has been shown to be highly dependent on the operating temperature and the workload executed by the circuit [2, 3]. Moreover, circuit reliability is also affected by process-induced device's variations (PV) [4, 5], which have a significant impact on circuit performance and make more difficult to satisfy stringent reliability constraints during circuit design.

The conventional approach to assure circuit lifetime under BTI and PV effects is to add a worst-case guardband to the clock period. In such a way, correct signal propagation through the logic paths is assured. However, as devices continue to shrink, the required guardbands are becoming

Responsible Editor: L. M. Bolzani Pöhls

✉ Andres Gomez
fgomez@inaoep.mx; andres.gomez@docentes.umb.edu.co

Victor Champac
champac@inaoep.mx

[1] National Institute for Astrophysics, Optics and Electronics (INAOE), Luis Enrique Erro #1, Tonantzintla, Puebla, Mexico

[2] Universidad Manuela Beltrán (UMB), Calle 33#27–12 Bucaramanga, Santander, Colombia

unacceptably large, leading to conservative designs with reduced performance [6, 7].

Various aging-aware design techniques already exist in the literature. In [8, 9], gate's input-node-reordering was proposed to mitigate the delay degradation of the paths due to BTI. The idea was to manipulate the percentage of time the devices experience BTI stress, also known as stress probability. However, degradation reduction may be insufficient to mitigate guardbands under both BTI and PV effects. Gate size optimization is a widely used approach to address aging and process variations issues. This method resizes the gates to achieve optimal trade-offs between delay, area, and lifetime reliability. In [10], the design optimization of a full-adder circuit based on extensive SPICE simulations was presented. However, SPICE-based optimization is computationally unfeasible for large-scale integrated circuits. The works [11, 12], built gate libraries robust to aging by sizing the transistors in a gate according to the stress probability that the devices experience. These approaches require more detailed guidance to determine where to place the robust gates within a circuit as the gates with the largest delay degradation may not be the most influential to overall circuit timing. In [13], it is proposed to increase the size of all the gates liying in the critical paths of the circuit and having a delay degradation larger than a given threshold (i.e., 5%). The proposed approach takes into account the maximal load capacitance that a gate of a given size can drive. However, not all the gates in the critical paths have the same impact on circuit delay degradation. Therefore, they should be treated differently. In [14, 15] an optimization problem minimizing circuit area for a given delay constraint is formulated and solved using Lagrangian relaxation. These methods may become complex for large circuits, especially if process parameters variations are considered.

The concept of gate criticality metrics under aging effects was introduced in [16]. Gate criticality metrics provide a fast estimation of how efficiently the delay degradation of the circuit improves at a given area or power cost when sizing a gate. Then, design actions can take place based on the metric scores. Different gate criticality metrics have been proposed in [5, 16, 17], and [18]. In [16, 17], the selected gates are replaced by their aging-robust counterparts from an aging-aware gate library (such as that in [12]). In [5, 18], the size of the gates with the highest metric score is iteratively increased until the desired timing constraint is met. However, it is not considered to decrease the size of gates with little impact on delay to mitigate area overhead. Also, the used metrics do not consider the impact of sizing a gate on both the degradation and the standard deviation of the paths delay (under PV), which may limit the efficiency of the optimization process.

Aging-aware circuit design optimization becomes a complex problem in scaled technologies because BTI-induced delay degradation strongly depends on the execcuted workload, which defines the stress probability of each transistor in the circuit. Unfortunately, the exact workload executed by a circuit over the lifetime is unpredictable and hardly to know in advance at the design phase. Therefore, a major limitation of the aforementioned aging-aware optimization approaches is that they either assume worst-case stress probability or a specific signal proability profile at main circuit inputs for aging estimation. While the first approach leads to conservative designs with excessive area overhead, the second approach may not be reliable if the actual signal probabilities of the circuit differ from those used during circuit design. Recently, a sizing approach considering the distribution of paths delay degradation for various workload profiles was proposed in [19]. The circuit is optimized based on the mean value of the delay degradation of the paths over a set of workloads, but this does not guarantee reliable operation. Furthermore, the effect of process variations was not considered.

This paper presents a methodology for guardband reduction by efficient selection and sizing of critical gates considering BTI aging and PV effects. This is an extension of our previous work in [20]. The proposed approach uses metrics to identify those gates providing efficient guardband reduction with as small as possible area overhead. The main contributions of this paper are:

1. A multiple workload-aware sizing algorithm is proposed. The paths delays are estimated for various workload scenarios at main inputs. In such way, a more accurate estimation of the maximal paths delay degradation is made. Then, the paths are optimized for the workload scenario that causes the largest delay degradation.
2. New statistical gate sizing metrics are proposed. The metrics include the impact of gate sizing on the BTI delay degradation and the standard deviation of the delay. A fast approximation for the sensitivity of the statistical delay of a path with respect to the size of a gate is proposed. The optimization process considers sizing-up gates to improve delay and sizing-down gates to mitigate area overhead.

The rest of this paper is organized as follows: Section 2 explains path-based delay estimation under BTI-aging and Process Variations. Section 3 presents the proposed gate size optimization methodology. Section 4 presents the proposed metrics and the sizing heuristic for guardband reduction with low area cost. Section 5 presents the simulation results on ISCAS Benchmark circuits. Section 6 presents the conclusions of this work.

2 Delay Estimation Under BTI and Process Variations

2.1 Statistical Model for BTI Aging

Bias Temperature Instability (BTI) is the dominant aging mechanism in modern technologies. Negative-BTI (NBTI) affects PMOS transistors under a negative gate-to-source bias. Similarly, Positive-BTI (PBTI) affects NMOS transistors under positive gate-to-source bias. NBTI was considered the major reliability issue before the 45nm technology node. However, PBTI has become important since the introduction of the high-k metal gate dielectric in sub-45nm technologies [21]. BTI mechanism has two phases [22, 23]:

1. Stress Phase: BTI is associated with the degradation of the $Si - SiO_2$ interface of the device due to the breaking of weak $Si - H$ bonds caused by the high vertical electric field and elevated temperatures. The released H atoms combine to form H_2 spices and diffuse into the oxide leaving an interface-trap [22]. BTI is also associated with the trapping and de-trapping of charge carriers from the channel tunneling into pre-existing traps (defects) in the gate oxide [23]. These mechanisms manifest as a gradual increase on devices V_{th} during the stress phase.
2. Recovery Phase: When stress is removed ($|Vgs| = 0$) some of the traps in the $Si - SiO_2$ interface are passivated. Therefore, the V_{th} degradation during the stress phase is partially recovered.

The overall increase in V_{th} is a function of the percentage of time the device is at stress, also known as the stress probability, which strongly depends on the executed workload by the circuit. A power law is widely accepted to model this dependence [24–26]. A closed form equation to calculate BTI-induced V_{th} degradation is [26],

$$\Delta V_{th,BTI} \approx K \cdot t_{ox} \cdot \sqrt{C_{ox} \cdot (V_{GS} - V_{TH0})} \cdot e^{\left(\frac{E_{ox}}{E_0}\right)} \cdot e^{\left(\frac{-E_a}{kT}\right)} \cdot \alpha^n \cdot t^n \quad (1)$$

where n is the time exponent, t_{ox} is the gate oxide thickness, E_{ox} is the vertical electric field, T is the temperature, k is the Boltzmann constant, C_{ox} is the oxide capacitance per unit of area, V_{TH0} is the initial (fresh) threshold voltage value, E_a and E_0 are constants, α is the stress probability and K is a technology-dependent fitted constant, which can be different for NBTI and PBTI.

As can be observed in Eq. 1, the V_{th} deterioration depends on the initial V_{th} (V_{TH0}). However, V_{TH0} becomes a random variable due to process variations. The impact of process variations in the long-term degradation of V_{th} can be accounted by a first-order Taylor approximation of Eq. 1 [27],

$$\Delta V_{th,BTI} = (1 + S_v \cdot \Delta V_{th,PV}) \cdot A \cdot \alpha^n \cdot t^n \quad (2)$$

where $\Delta V_{th,PV}$ is the shift in V_{TH0} due to process variations, and A and S_v are fitted constants. Then, the total V_{th} variation of a transistor m corresponds to the summation of the contributions due to BTI ($\Delta V_{th,BTI}$) and Process variations ($\Delta V_{th,PV}$), as given by Eq. 3 [27],

$$\Delta V_{th,m} = A_m \cdot \alpha_m^n \cdot t^n + (1 + S_{V,m} \cdot A_m \cdot \alpha_m^n \cdot t^n) \cdot \Delta V_{th,PV,m} \quad (3)$$

Note that at the beginning of the lifetime ($t = 0$) the total variation in V_{th} is due to only process variations. However, as circuit ages, BTI causes a shift in both the mean value and the variance of V_{th} [28].

2.2 Aging-Aware Statistical Gate Delay Model

For Statistical Static Timing Analysis, the gate delay is modeled as a linear function of normally distributed random variables representing process parameters.

$$D = D_n + S_W^D \Delta W + S_L^D \Delta L + S_{tox}^D \Delta t_{ox} + \sum_m^M S_{V_{th,m}}^D \Delta V_{th,m} \quad (4)$$

where D_n is the nominal gate delay, S_W^D, S_L^D, S_{tox}^D and $S_{V_{th}}^D$ are the gate delay sensitivities with respect to deviations in W, L, tox, and V_{th}, respectively. M is the number of transistors in the gate. Note that ΔV_{th} is composed of two deviation components, one related to the time-zero variability and the other related to aging effects (See Eq. 3). This linear model is adequate for small enough variations as computational complexity remains low and the error due to discarded higher order terms can be neglected [29].

In order to use the Aging-Aware Statistical Gate Delay Model into a Statistical Static Timing Analysis tool, the parameters in Eq. 4 are pre-computed by accurate SPICE electrical simulations. For each gate type (i.e., INV, NANDs, NORs), HSPICE simulations are run at various design conditions given by combinations of the input transition time (SRIN), the gate size (K), load capacitance (CL), and the operating Temperature (Te). For each combination, the nominal gate delay and gate delay sensitivities to process parameters are measured. Then, the extracted data is fitted using polynomials, which allow a fast and accurate estimation of the statistical gate delays using Eq. 4.

2.3 Statistical Delay of a Path

The statistical delay of a path is computed as the statistical sum of the random variables representing the delay of each

gate in the path. Given the mean and standard deviation for a given aging time for all the gates in the path, the PDF of the path ($D_p = N(\mu_{D,p}, \sigma_{D,p})$) is obtained by:

$$\mu_{D,p} = \mu_{Dn,p} + \mu_{\Delta D,p} = \sum_{i=1}^{N} \mu_{Di} \tag{5a}$$

$$\sigma_{D,p} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} \rho_{ij} \cdot \sigma_{Di} \cdot \sigma_{Dj}} \tag{5b}$$

where $\mu_{Dn,p}$ is the mean of the nominal delay of the path, $\mu_{\Delta D,p}$ is the mean of the delay degradation of the path, μ_{Di} is the mean of the aged delay of the gate i for the given aging time, σ_{Di} and σ_{Dj} are the standard deviation of the aged delay of gates i and j, respectively. The parameter ρ_{ij} is the correlation between gate delays, which depends on the spatial proximity of the gates in the circuit layout. The analytical model proposed in [30] is used to estimate the degree of spatial correlation between two gates. Note that the mean delay value of a path has a nominal component ($\mu_{Dn,p}$) and a component due to aging effects ($\mu_{\Delta D,p}$). Also note that the standard deviation of the delay of a path depends on aging effects, as the threshold voltage variability changes due to aging (See Eq. 3).

3 Proposed Methodology for Guardband Reduction by Selection and Sizing of Critical Gates

The proposed optimization methodology consists of the three steps shown in Fig. 1. In the first step, those paths that may become critical under worst BTI conditions (worst stress probability and worst temperature) are identified. Those paths are called the *Potential Critical Paths* (PCPs) of the circuit. Similarly, the gates belonging to these paths are called the *Critical Gates* of the circuit. In this paper, only the PCPs are considered during design optimization. The non-PCPs are not considered for optimization as they would not trigger any aging-related issue.

In the second-step, a multiple work-load-aware aging analysis of the PCP set is done to estimate the specific workload that causes a realistic maximum aged delay on each PCP. In the third step, the PCPs are optimized using the proposed gate sizing metrics so that their realistic maximum aged delay satisfy a given target guardband (GB_t) with low area cost.

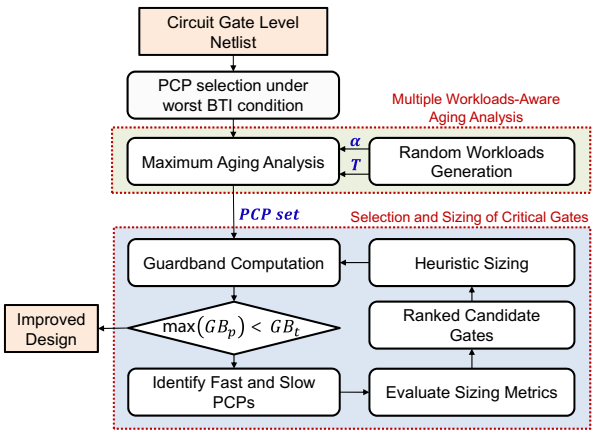


Fig. 1 Flow of the proposed gate sizing optimization methodology

3.1 PCP Identification Under Worst BTI Condition

Aging-Aware Statistical Static Timing Analysis (SSTA) is run assuming worst BTI conditions, i.e., the devices in the circuit are assumed to operate under near-static stress ($\alpha \approx 1$) and high temperature ($T = 120°C$). Those paths with a $\mu + 3\sigma$ of the aged delay distribution greater than the nominal (without aging and PV) delay of the circuit are identified as Potential Critical Paths (PCPs).

The identification of PCPs under worst BTI conditions allows focusing the optimization in a reduced path set rather than in the entire circuit, reducing computational effort.

3.2 Multiple workload-Aware Aging Analysis

A workload corresponds to the set of consecutive bits applied to each main input of the circuit when executing a given program [31] and it is represented by the Signal Probability (SP) at main circuit inputs (probability of a node to be at logic 1). The workload impacts the stress probability (α) of each device and on their operating temperature [2, 3], which in turn influence BTI degradation, making complex circuit reliability analysis and optimization.

To address the unpredictability of the circuit workload at the design phase, we refine the workload conditions at which the delay of each PCP is evaluated during design optimization by performing a *Multiple Workload-Aware Aging Analysis*. The idea behind this step is to determine the workload at which a realistic maximum delay degradation of each PCP occurs. Figure 2 shows a histogram of the mean of the aged delay of a PCP in ISCAS circuit c2670 for 1000 different workload profiles. As can be seen, the maximum aged delay that the path can take over all tested workload profiles is much lower than the aged delay estimated using worst BTI conditions ($\alpha \approx 1$ and $T = 120°C$). This is because the devices under the tested workload profiles experience more realistic degradation conditions due to BTI. Figure 2 also shows that the variation of path delay degradation due to the workload can be approximated by a gaussian-like distribution, as was also found in [3, 19].

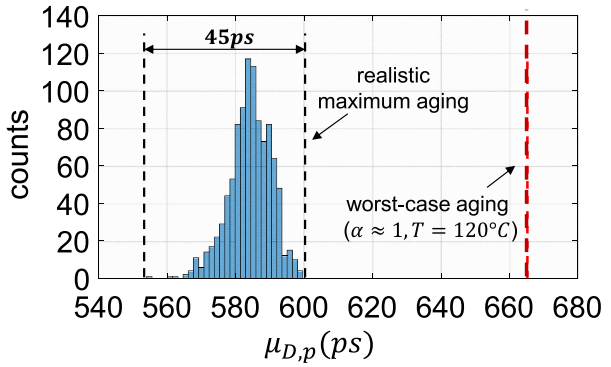


Fig. 2 Histogram of the delay of a PCP for various workload profiles (ISCAS c2670)

Since it is unfeasible to evaluate the delay degradation for each path and for every possible combination of signal probabilities at main inputs (representing a workload profile) for a state-of-art digital circuit, the following strategies are proposed to estimate an upper bound for the delay degradation of the paths with an acceptable computational cost:

- The multiple workload-aware aging analysis is only performed over the PCP set.
- For each PCP, only its mean delay degradation due to process variations is computed for the tested workloads.
- If the delay degradation being obtained for a PCP does not increase after testing a given number N of consecutive workload profiles, it is assumed that a good enough approximation of the maximum PCP aged delay has been obtained, and the PCP degradation is not longer computed for the subsequent workload profiles.
- Once the workload that causes maximum delay degradation for each PCP is identified, SSTA is run to compute the deviation of the delay of the PCPs due to process variations.

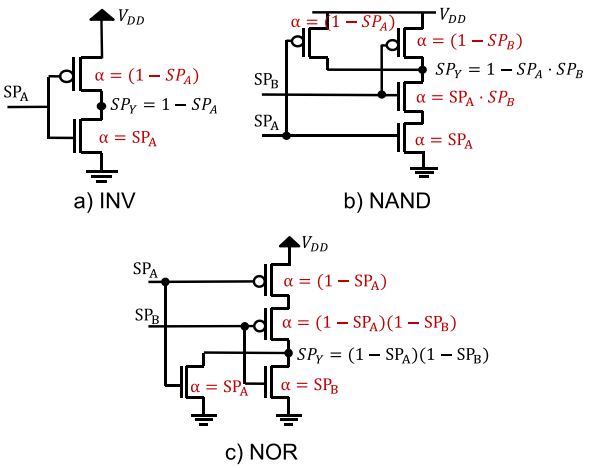


Fig. 3 Signal Probability Propagation and Stress Probability computation rules

Algorithm 1 Multiple workload-aware maximum aging analysis.

Input: PCP set, $MaxWL$
Output: α and T of PCP gates causing largest aging

```
1:  for WL = 1 to WL = MaxWL do
2:      generate_propagate_SP()
3:      compute_stress_probability()
4:      compute_temperature()
5:      compute_ΔV_th,BTI()
6:      for p = 1 to p = Num.PCPs do
7:          if PCP[p].MAX == 0 then
8:              Compute μ_D,p
9:              if μ_D,p is the current maximum delay for p
                then
10:                 Save α and T of each device in PCP p
11:             else
12:                 if a larger μ_D,p was not obtained in the
                    past N workloads then
13:                     PCP[p].MAX=1 (p is not evaluated
                        for next WL)
14:                 end if
15:             end if
16:         end if
17:     end for
18: end for
```

Algorithm 1 describes the proposed multiple workload-aware aging analysis procedure. For an user-defined number of workload profiles ($MaxWL$), a set of signal probabilities at main circuit inputs are generated and propagated to internal nodes (function ***generate_propagate_SP()***). A uniform random number generator between 0 and 1 is used to obtain the signal probability assigned to each input. Then, the stress probability (α) of each transistor in the circuit is computed (function ***compute_stress_probability()***). Figure 3 illustrates the basic equations for signal probability propagation and stress probability computation for some basic gates. The formula to propagate the signal probabilities for other more complex gates can be easily derived based on their truth tables. The operating temperature of each cell is also computed as it strongly influences BTI mechanism (function ***compute_temperature()***). The temperature profile of the circuit is obtained from the power consumption profile using the electric model given in [32],

$$T_i = R_{J,i} \cdot P_i + R_{I-A} \cdot P_{total} + T_A \quad (6)$$

where T_i is the operating temperature of gate i, P_i is the power consumption (Static and Dynamic) of the gate i, $R_{J,i}$

is the junction to internal air heat resistance, P_{total} is the total circuit power consumption, R_{I-A} is the heat resistance from internal air to ambient, and T_A is the ambient temperature [32]. Once the stress probability and operating temperature are obtained, the BTI-induced V_{th} shift of each device is computed (function $\boldsymbol{compute\_\Delta V_{th,BTI}()}$). Then, the mean value of the aged delay ($\mu_{D,p}$) is computed for each PCP p whose flag variable $PCP[p].MAX$, which indicates that a *good enough* maximum aged delay of the path has been found, is not activated. If the obtained $\mu_{D,p}$ is the largest obtained for the currently tested workloads, the conditions of stress probability and temperature of the devices in the path are stored. If the obtained $\mu_{D,p}$ is not larger than the previous $\mu_{D,p}$ computed for a consecutive user-defined number (N) of workload profiles, the flag variable ($PCP[j].MAX$) is activated, indicating that the currently stored conditions for the path p cause a *good enough* estimation of the maximum aged delay of the path. Then, this path is not evaluated for the subsequent workload profiles. It is important to note that the workload that causes maximum path delay degradation can be different for each path.

Once the workload condition that causes maximum delay degradation for each PCP is identified, SSTA is run to compute the standard deviation of the delay of the PCPs. Then, the set of PCPs is reduced by discarding those paths whose maximum aged delay at the $\mu + 3\sigma$ corner does not exceed the nominal circuit delay. This process mitigates the computational effort required for design optimization. Moreover, the corresponding workload condition that causes a maximum delay degradation for each PCP is stored so that the path delay can be re-evaluated under such conditions if needed.

Figure 4 shows the behavior of the cumulative maximum delay degradation obtained for some paths of the circuit C1908 as a function of the number of tested workloads. As can be seen, the maximum delay degradation obtained for all the paths tend to saturate after some workload profiles are tested. This behavior suggests that only a moderated number of workload profiles need to be analyzed to get a good estimation of the maximum aged delay that a path can take.

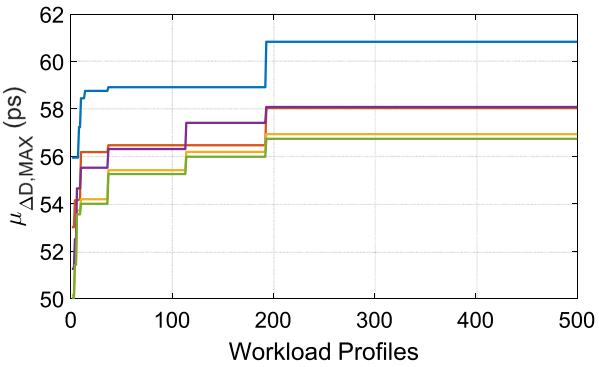


Fig. 4 Maximum delay degradation of some paths as function of the number of tested workload profiles

4 Selection and Sizing of Critical Gates

This section presents the proposed methodology for selection and sizing of the critical gates to optimize the circuit to satisfy a reduced target guardband (GB_t).

4.1 Guardband Computation

The first step for the selection and sizing of critical gates (See Fig. 1) is to compute the actual guardband of the circuit. Here, only the maximum aged delay of each PCP that was obtained from the multiple workload-aware aging analysis step is considered. The guardband that each PCP impose (GB_p) over the nominal circuit delay is defined as,

$$GB_p = (\mu_{D,p} + 3\sigma_{D,p}) - D_{nom} \tag{7}$$

where μ_{Dp} and $\sigma_{D,p}$ are the mean value and the standard deviation of the maximum aged delay of the PCP p, and D_{nom} is the nominal circuit delay (no BTI and no PV).

The proposed methodology in this work assures reliable circuit operation for a user defined Target Guardband (GB_t), which is smaller than the Initial Guardband, under the combined effect of aging and process variations.

4.2 Identification of Fast and Slow PCPs

The PCPs are then separated into two different subsets depending on the corresponding guardband imposed by each path, as illustrated in Fig. 5a) Slow-PCPs subset, which has negative slack ($GB_t - GB_p < 0$); and b) Fast-PCPs subset, which has positive slack ($GB_t - GB_p > 0$). This classification is done to exploit the fact that different design actions can be taken over each PCP subset. Some gates in the Slow-PCPs are sized-up to improve their delay,

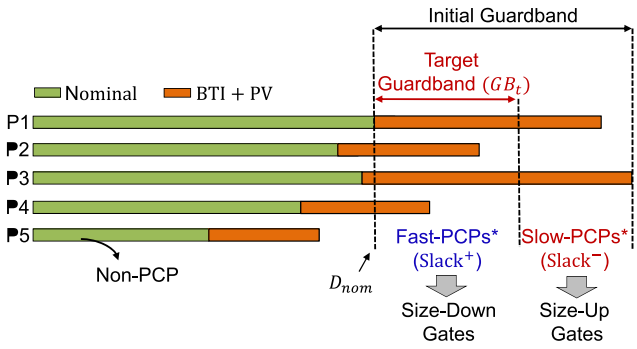


Fig. 5 Fast and Slow PCP sets

while some gates in the Fast-PCPs are sized-down to take advantage of their slack to mitigate area overhead.

4.3 Evaluation of Sizing Metrics

Gate selection metrics are proposed to guide the optimization process. The metrics are intended to identify the best critical gates to be sized in each PCP subset to efficiently improve the circuit guardband.

4.3.1 Sensitivity of the Statistical Delay of a Path to a Gate Size

We define the sensitivity of the statistical delay of a path with respect to the size of a gate as the derivative of the $\mu + 3\sigma$ of the path delay distribution to a change in the size of the gate i in the path:

$$S_{Ki}^{Dp} = \frac{\partial \mu_{Dp}}{\partial K_i} + 3 \cdot \frac{\partial \sigma_{Dp}}{\partial K_i} = \left[\frac{\partial \mu_{Dn,p}}{\partial K_i} + \frac{\partial \mu_{\Delta D,p}}{\partial K_i}\right] + 3 \cdot \frac{\partial \sigma_{Dp}}{\partial K_i} \quad (8)$$

where K_i is the size of the gate i in the path, μ_{Dp} and σ_{Dp} are the mean value and the standard deviation of the aged path delay obtained with Eqs. 5a and 5b, respectively. $\mu_{Dn,p}$ and $\mu_{\Delta D,p}$ correspond to the mean value of the nominal (fresh) path delay and the mean value of the delay degradation of the path.

Equation 8 measures the impact of sizing a gate on the path delay. As can be seen, three components influence S_{Ki}^{Dp}: 1) the component related to the nominal delay (no aging and no PV), 2) the component related to aging effects, and 3) the component related to process variations. Figure 6 shows these components for the path example shown in the inset Figure. As can be observed, the component related to the nominal path delay is the largest. However, the components due to the impact of aging on the mean delay and the impact of process variations are also important. It is worth to mention that the aging component depends on the degradation of the gate. A gate whose devices have larger aging also exhibit a larger $\frac{\partial \mu_{\Delta D,p}}{\partial K_i}$. It is also important to note that spatial correlation plays an important role in the magnitude of $\frac{\partial \sigma_{Dp}}{\partial K_i}$. Figure 6 shows two cases: when all the gates in the path are placed far away, and their spatial correlation is almost zero ($\rho = 0$), and the case when all the gates are placed very close to each other, having a full spatial correlation ($\rho = 1$). Therefore, those gates that have a higher correlation with the other gates in the path may be preferable to be optimized.

The brute-force approach for computing Eq. 8 is to evaluate the statistical distribution of the aged path delay for both the current size of the gate and when the size of the gate is changed by a small perturbation (this is done for the numerical computation of the derivatives). In such way, for a path with N gates, the statistical delay of the path would have to be computed $N + 1$ times to compute the sensitivity of the statistical delay of the path with respect to the size of each gate, which is computationally costly. Therefore, we propose some simplifications to evaluate Eq. 8 more efficiently, as explained next.

Figure 7b shows the derivative of the mean value and the standard deviation of the delay of each gate in the path shown in Fig. 7a to a change in the size of the gate i in the path. As can be seen, only the timing response of the gates $i-1$, i, and $i+1$ are significantly affected. We call the set of these gates as the *path segment* for gate i. As shown, both the mean and standard deviation of the gate $i-1$ increases

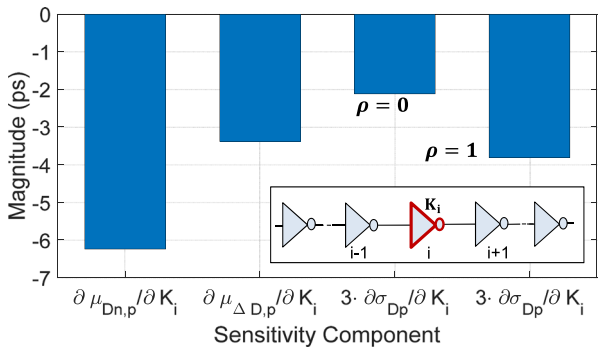
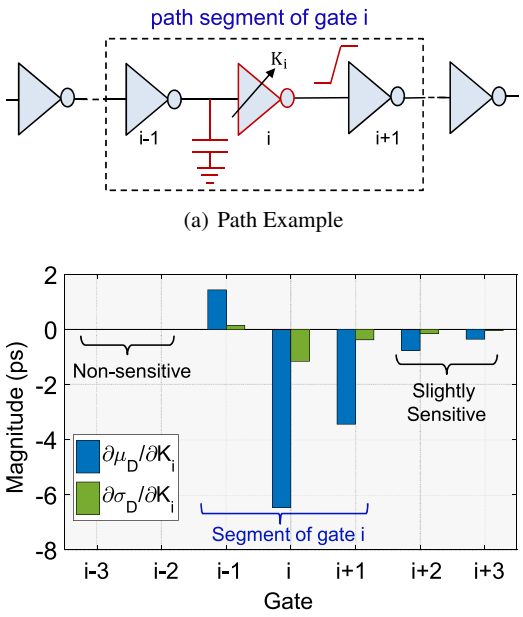


Fig. 6 Example of the magnitude of the components of the sensitivity of the statistical delay of a path to sizing of a gate (Eq. 8)



(a) Path Example

(b) Derivative of the mean value and standard deviation of the delay of each gate in the path to the size of gate i.

Fig. 7 A path example to illustrate the impact of sizing a gate on its neighboring gates in the path

due to the larger input capacitance of the sized gate. On the other hand, the mean and standard deviation of the gate $i+1$ reduces because its input signal switches faster as gate i becomes stronger. Obviously, the mean value and the standard deviation of the delay of the sized gate are the most reduced when the size of this gate is increased. It should be noted that the change in the standard deviation of the delay of a gate is much smaller than the change in the mean value, as was observed before in Fig. 6. Based on the above mentioned observations, the following approximations are made:

Sensitivity of the Mean of the Path Delay to Gate Sizing It is assumed that a change in the mean delay of a path is mainly due to a change in the mean delay of the gates in the *path segment* of the gate i. Therefore, we approximate $\frac{\partial \mu_{Dp}}{\partial K_i}$ as,

$$\begin{aligned}\frac{\partial \mu_{Dp}}{\partial K_i} &\approx \frac{\partial \mu_{D,i-1}}{\partial K_i} + \frac{\partial \mu_{D,i}}{\partial K_i} + \frac{\partial \mu_{D,i+1}}{\partial K_i} \\ &\approx \frac{\partial \mu_{D,i-1}}{\partial CL_{i-1}} \cdot \frac{\partial C_{in,i}}{\partial K_i} + \frac{\partial \mu_{D,i}}{\partial K_i} + \frac{\partial \mu_{D,i+1}}{\partial SRI_{i+1}} \cdot \frac{\partial SRO_i}{\partial K_i}\end{aligned} \tag{9}$$

where $\mu_{D,i-1}$, $\mu_{D,i}$ and $\mu_{D,i+1}$ are the aged delays of the gates $i-1$, i and $i+1$ in the path segment of the gate being analyzed, CL_{i-1} is the load capacitance of the gate $i-1$, $C_{in,i}$ is the input capacitance of gate i, SRI_{i+1} is the signal transition time at input of gate $i+1$ and SRO_i is the signal transition time at output of gate i, which is equal to SRI_{i+1}.

Note that by using this approximation only the mean delay of the path segment of the gate i needs to be recomputed.

Sensitivity of the Standard Deviation of the Path Delay to Gate Sizing It is assumed that the change in the standard deviation of the delay of a path due to sizing a gate i is mainly due to the change of the standard deviation of the delay of the gate i and its impact on the covariance with the other gates in the path. We can write:

$$\begin{aligned}\frac{\partial \sigma_{D,p}}{\partial K_i} &= \frac{1}{2\sqrt{\sigma^2_{D,p}}} \cdot \frac{\partial\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{ij}\cdot\sigma_{Di}\cdot\sigma_{Dj}\right]}{\partial K_i} \\ &\approx \frac{1}{2\sigma_{D,p}} \cdot \left(\frac{\partial \sigma^2_{Di}}{\partial K_i} + 2\sum_{j\neq i}^{N}\frac{\partial \sigma_{Di}}{\partial K_i}\cdot\rho_{ij}\cdot\sigma_{Dj}\right) \\ &\approx \frac{1}{\sigma_{D,p}} \cdot \left(\frac{\partial \sigma_{Di}}{\partial K_i}\sum_{j=1}^{N}\rho_{ij}\cdot\sigma_{Dj}\right)\end{aligned} \tag{10}$$

As can be observed, the sensitivity of the standard deviation of the path delay depends on the spatial correlation that the sized gate i has with each other of the gates in the path. Note that Eq. 10 only depends on the derivative of the standard deviation of the delay of the gate i with respect to the size of the gate itself. Therefore, to evaluate Eq. 10 only the standard deviation of the gate of interest i needs to be recomputed.

4.3.2 Proposed Gate Sizing Metrics

The statistical sensitivity S^{Dp}_{Ki} reveals which gate has a larger impact on the $\mu+3\sigma$ delay of the path. This parameter is combined with other important information of the gates to form the proposed gate sizing metrics.

Two gate sizing metrics are proposed to guide the optimization process: One that measures the benefit of sizing-up a gate in the Slow-PCPs, and other that measures the benefit of sizing-down a gate in the Fast-PCPs. For each gate i, the two following metrics (See Eq. 11) are evaluated:

$$M_{SU,i} = \frac{S^{D}_{Ki,AVG}\cdot|Slack^{-}_{i,AVG}|\cdot N_i}{\Delta A_i} \quad M_{SD,i} = \frac{Slack^{+}_{i,AVG}\cdot\Delta A_i}{S^{D}_{Ki,AVG}\cdot N_i} \tag{11}$$

where $M_{SU,i}$ and $M_{SD,i}$ are the sizing-up and sizing-down metrics, respectively. $S^{D}_{Ki,AVG}$ is the average statistical delay sensitivity of the N_i paths passing through the gate i with respect to changes in gate size (Ki), $Slack_{i,AVG}$ is the average slack of the paths passing through the gate i, and ΔA_i is the area impact of sizing the gate, which depends on the geometry of the cell layout. Note that each metric is evaluated for a different PCP set. For sizing-up metric, $Slack_{i,AVG}$ takes a negative value as it is evaluated over the Slow-PCP set. On the other hand, $Slack_{i,AVG}$ takes a positive value for the sizing-down metric, where the Fast-PCPs are considered (See Fig. 5). The value of N_i and $Slack_{i,AVG}$ are also different depending on the PCP set being considered.

The metric score determines the delay-area trade-off of sizing a gate. The sizing-up metric score increases for gates influencing many paths since they allow to improve various paths at a time. The sizing-up metric score also increases for those gates in Slow-PCPs with large negative slacks as those paths should be optimized with higher priority. A large average statistical path delay sensitivity with respect to gate sizing also increases the sizing-up metric score as a large delay reduction can be obtained by increasing the gate size. Finally, the sizing-up metric score reduces for gates with a high area impact because increasing the size of those gates is area costly. A similar interpretation of the parameters is made for the sizing-down metric. In this case, the size-down metric score increases for those gates affecting few Fast-PCPs with low delay sensitivity to gate sizing (low impact on delay) and large positive slack. Also, gates with a large

area impact are preferred due to potential area savings when sizing-down a gate.

4.4 Sizing Heuristic

Algorithm 2 summarizes the sizing heuristic.

The obtained sizing-up metric score $M_{SU,i}$ reflects the benefit of Slow-PCPs delay reduction vs. area trade-off of each gate. Thus, N gates with the highest $M_{SU,i}$ are picked and size-up proportionally to their respective score: $\Delta K = step \cdot M_{SU,i}$. Where N is an user-defined number of gates that are sized at each iteration and $step$ is the maximum size change that a gate can take at an iteration.

The sizing-down metric score $M_{SD,i}$ reflects a trade-off between the delay increase of the Fast-PCPs and the area reduction. However, the interdependence between Fast-PCPs and Slow-PCPs must be considered to select the gates to be sized-down because a gate having a high $M_{SD,i}$ may negatively impact on Slow-PCPs if the gate also has a high $M_{SU,i}$ score. Therefore, the two following conditions are applied to select the gates to be sized-down:

- Gates sized-up are not allowed to be sized-down in the same iteration.
- Gates that have a sizing-up metric score ($M_{SU,i}$) larger than a constraint ($C_{M_{SU}}$) are not allowed to be sized-down.

The constraint $C_{M_{SU}}$ is used to limit the negative impact on the slow-PCPs delay of sizing-down gates. The value of the constraint is dynamically changed along the sizing process. It is initially set to 1 (maximum) to maximize area savings as any gate is allowed to be sized-down, but it is gradually reduced each time the delay of the Slow-PCPs is not improved in a given iteration, so that the guardband converges towards the desired target delay. The N gates with the highest $M_{SD,i}$ score fulfilling the aforementioned conditions are sized-down according to the following rule: $\Delta K = -step \cdot (1 - M_{SU,i}) \cdot M_{SD,i}$. Thus, the amount of size reduction of a selected gate reduces (increases) if the gate has a high (low) $M_{SU,i}$ ($M_{SD,i}$) score.

The size-down procedure is useful when the initial design has oversized gates due to a non-optimal design. Also, it becomes beneficial when a gate in a Fast-PCP is driven by a gate in a Slow-PCP. This may occur if the gate in the Fast-PCP was sized-up at the beginning of the optimization procedure (i.e., the gate was critical first), but then its importance to the remaining Slow-PCPs decreases.

Once the selected gates are sized, the PCPs timing information is updated (See Algorithm 2) under the conditions of temperature and stress probability of the devices that caused maximum aged path delay, obtained from the multiple workload-aware aging analysis steps.

Algorithm 2 Sizing heuristic.

Input: Gates Metrics Scores ($M_{SU,i}$ and $M_{SD,i}$)
Output: Selected Gates with Updated Size

```
Set C_MSU = 1
Rank gates in Slow-PCPs according to M_SD,i
Rank gates in Fast-PCPs according to M_SD,i // Size-up gates in Slow-PCPs:
for i = 1 to N do
    K_i = K_i + step · M_SU,i
end for// Size-Down gates in Fast-PCPs
for i = 1 to N do
    if M_SU,i < C_MSU and i was not sized-up then
        K_i = K_i − step · (1 − M_SU,i) · M_SD,i
    end if
end for
Update PCPs Timing Information
if max(GB_p) is not reduced then
    C_MSU=C_MSU − 0.1
end if
```

5 Simulation Results on ISCAS Benchmark Circuits

The proposed gate sizing optimization technique for guardband reduction has been implemented in C ++ code and applied to ISCAS benchmark circuits designed using a 32nm Synopsys Generic Technology [33]. The original design of each circuit is of minimum area, where all the gates have minimum dimensions.

5.1 Statistical Path Delay Sensitivity Approximation

Let us first analyze the accuracy of the proposed approximation for the sensitivity of the statistical delay of a path. For this analysis, the impact of sizing each gate at the $\mu + 3\sigma$ delay of the slowest path of ISCAS circuit C1908 was computed. Figure 8 shows the sensitivity of the statistical delay of the path with respect to the size of each gate in the path. Data is shown for both the statistical path delay sensitivity obtained with the proposed derivative approximation (See Eqs. 8, 9 and 10) and the exact derivative calculation, where the statistical delay of the path is re-computed when the size of each gate in the path is perturbed. As can be observed, the proposed approximation follows well the derivative obtained with the exact computation.

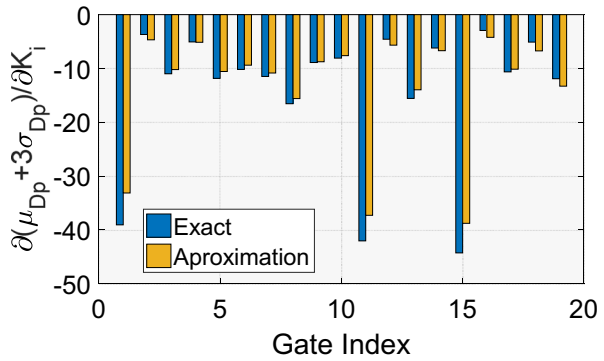


Fig. 8 Statistical delay sensitivity of a Path with respect to sizing each gate in the path. Longest path of C1908 circuit

5.2 Optimization Results

Table 1 shows detailed results obtained from the application of the proposed design optimization methodology to ISCAS 85/89 circuits. Circuits of different size and complexity were considered. The second and third columns give the total number of paths and gates in the circuits. Columns 4-7 show results related to the multiple workload-aware aging analysis step. The column labeled as PCPs correspond to the number of Potential Critical Paths, which are those paths whose $\mu + 3\sigma$ delay may become greater than the nominal delay of the circuit. These paths are the ones considered during selection and sizing of the gates. As can be observed, the number of PCPs does not depend on the total number of paths (i.e., the number of PCPs in c7552 and s1423 is very different, but these circuits have a similar number of paths). The number of PCPs changes depending on the susceptibility of each circuit to aging and the circuit topology. Column 5 gives the number of gates belonging to the selected set of PCPs. These gates are called as *Critical Gates* (CGs). The proposed heuristic uses the sizing metrics to identify which critical gates are more beneficial to be sized. Column 6 shows the initial guardband that would have to be added to the nominal delay to assure reliable circuit operation under the combined effect of aging and process variations. As can be seen, the percentage of guardband needed can be up to 45% of the nominal delay, which may be unacceptably large for high-performance state-of-art designs. Column 7 shows the CPU time spent in the multiple workload-aware aging analysis. This corresponds to the time for evaluating the PCP set for multiple workload profiles. It should be noted that the number of times each path is evaluated may be different depending on when it is detected that the maximal delay obtained for a path does not further increase when more workload profiles are analyzed.

Columns 8 to 13 of Table 1 show the results obtained applying the proposed methodology for selection and sizing of critical gates to reduce the initial guardband to a more acceptable target guardband of 20% (less stringent) and 10% (more stringent). The number of PCPs in the initial design that violate the corresponding target guardband (Slow-PCPs), the area overhead, and the CPU time for design improvement are given. When the guardband constraint is of 20% the area overhead for most of the circuits remains low because only some slow-PCPs out of the whole PCP set need to be improved. However, when the target guardband becomes more stringent, the number of slow-PCPs significantly increases for most of the circuits, depending on how balanced are the delays of the PCPs. The area overhead and the corresponding CPU time also increase for more stringet target guardbands as further optimization is needed to achieve the target.

5.3 Benefits of the Multiple Workload-Aware Aging Analysis

Tables 2 and 3 show the results for the cases when only one single workload and when worst BTI conditions are

Table 1 Optimization results using multiple workload-aware aging analysis

Circuit	Paths	Gates	Multiple Workload Analysis				Sizing: $GB_t = 20\%$			Sizing: $GB_t = 10\%$		
			PCPs	CGs	$GB(\%)$	$CPU\ (sec)$	Slow-PCP	$Area\ (\%)$	$CPU\ (sec)$	Slow-PCP	$Area\ \%$	$CPU\ (sec)$
c880	4935	254	1607	149	45.79	86.84	480	15.15	65.09	971	52.03	146.07
c1908	15638	253	8523	198	40.77	342.02	1727	9.76	271.73	4969	35.97	1386.80
c2670	3490	419	650	110	36.43	49.93	156	4.09	38.74	402	16.01	85.31
c5315	24666	1224	4785	403	35.77	302.50	677	2.38	204.81	2140	8.94	519.201
c7552	43613	1450	8070	935	38.25	442.08	781	0.32	143.53	3401	1.22	257.65
s298	231	166	79	36	40.44	2.57	23	6.26	1.17	46	16.34	2.52
s838	1714	279	262	102	38.09	13.73	73	4.92	3.066	150	11.01	6.62
s1423	44726	991	4323	339	30.75	1285.01	183	1.25	608.43	1180	4.91	1295.85
s5378	11728	1297	757	147	42.47	44.02	105	1.13	15.85	422	4.72	43.51

Table 2 Results using a single workload for aging analysis

Circuit	PCPs	CGs	$GB(\%)$	Sizing: $GB_t = 20\%$			Sizing: $GB_t = 10\%$		
				slow-PCP	A	CPU	slow-PCP	A	CPU
c880	1399	137	41.99	359	10.91	47.78	791	35.77	96.993
c1908	8102	196	39.42	1370	7.14	228.39	4322	30.85	919.49
c2670	630	106	35.41	143	3.47	35.16	372	14.55	90.93
c5315	4641	403	35.51	637	2.16	187.42	2066	8.63	617.04
c7552	7752	927	36.61	622	0.27	120.83	3047	1.07	247.36
s298	73	36	39.8	22	5.70	1.02	45	15.39	2.49
s838	258	102	37.18	67	4.23	2.59	139	10.18	6.152
s1423	4144	339	30.56	166	1.00	522.07	1110	4.90	1252.60
s5378	735	140	42.03	88	0.99	12.59	376	4.25	41.19

assumed for aging analysis, respectively. When only a single workload profile is used, the number of PCPs, the number of Critical Gates and the estimated guardband for the circuit are smaller than those obtained with our proposed multiple workload-aware aging analysis approach. This is because, in our approach, at least one of the tested workload profiles caused more aging in the PCPs than the workload profile assumed for the single workload case. Consequently, the area overhead when designing circuits using the single workload assumption is lower than the area overhead obtained with our proposal. Also, the CPU time for design optimization is slightly lower. However, if the workload profile that the optimized circuit experiences over the lifetime is different than the one used at design, some of the paths may degrade enough to cause a failure to time specifications. When only the worst BTI is assumed (See Table 3), the number of PCPs, Critical Gates and estimated GB for the circuit significantly increases, which results in significant area overhead and extra CPU time since the PCPs required more sizing than needed. For instance, consider circuit c2670, where 12.68% of the area overhead is saved when using our approach with respect to the design using worst-BTI conditions for a target guardband of 20%. The saved area increases to 49.19% for a stringent target guardband of 10%. A similar observation can be made for the other circuits.

The robustness of the optimized designs for 20000 random generated workload profiles was analyzed. For each workload profile, the corresponding stress probability and operating temperature of the devices were computed, and SSTA was performed to obtain the corresponding $\mu + 3\sigma$ delay of all the PCPs. Then, the maximum $\mu + 3\sigma$ delay among all the PCPs was identified, since this value corresponds to the maximum delay that the circuit can take for the given signal probability profile. Figure 9 shows histograms of the $\mu + 3\sigma$ delay of circuit s298 for both the optimized design ($GB_t = 10\%$) using our proposed multiple workload-aware aging analysis and the optimized design using only one single workload profile for aging analysis. As can be seen, there are some workloads for which the $\mu + 3\sigma$ delay of the circuit may violate the allowed 10% of guardband. However, it is clear that the optimized design with the proposed approach may violate the guardband for a significantly lower number of

Table 3 Results using worst BTI condition ($\alpha and T$) for aging analysis

Circuit	PCPs	CGs	$GB(\%)$	Sizing: $GB_t = 20\%$			Sizing: $GB_t = 10\%$		
				slow-PCP	A	CPU	slow-PCP	A	CPU
c880	2360	158	59.96	1004	54.33	210.47	1582	330.11	1526.04
c1908	11276	204	52.75	5409	38.12	4793.69	8610	125.90	4871.58
c2670	837	131	51.52	461	16.67	99.286	659	65.20	256.52
c5315	9331	551	53.22	3007	11.43	788.71	5656	44.57	2250.963
c7552	12866	1070	53.87	3615	1.30	449.40	7635	3.96	765.07
s298	88	38	55.18	54	21.95	3.28	76	127.44	26.72
s838	501	135	68.67	251	27.62	23.96	375	92.09	76.55
s1423	10885	408	48.13	2054	6.39	2973.07	5568	16.04	5162.69
s5378	1287	264	55.96	512	6.88	82.18	809	56.85	673.32

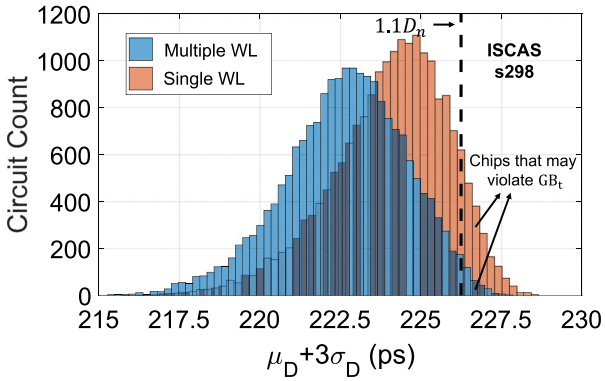


Fig. 9 Histograms of the $\mu + 3\sigma$ corner of the circuit aged delay obtained for an exhaustive number (20000) of multiple signal probability groups at circuit main inputs

workloads, which demonstrates the benefit of the proposed approach. Table 4 shows the percentage of workloads for which the $\mu + 3\sigma$ delay of the circuits violated the specified guardband constraint of 10%. For most of the circuits, the robustness of the optimized design using the multiple workload-aware aging analysis is significantly better than the optimized designs using only one single workload. Therefore, the obtained designs with the proposed approach are more reliable.

In the case that the coverage of possible workloads wants to be improved, designers can trade-off the degree of circuit reliability and the computational cost of performing a more exhaustive workload-aware aging analysis step.

5.4 Gate Sizing Optimization Comparison

The efficiency of the proposed gate sizing optimization metrics and the heuristic was compared against other aging-aware metrics proposed in [16] and [18], which are given in Eq. 12

$$M_{i,[16]} = \frac{N_i \cdot \Delta D_i}{max(N_i \cdot \Delta D_i)} + \delta \qquad M_{i,[18]} = S_{Ki}^{Di} \cdot \sum_p^N \Delta D_i \tag{12}$$

where $M_{i,[16]}$ is the metric proposed in [16], N_i is the number of paths (PCPs) passing through the gate i, ΔD_i is the delay degradation of the gate, and δ is a parameter that takes the value of 1 if the gate is in the slowest path of the circuit (the path with the largest negative slack). $M_{i,[18]}$ is the metric in [18], S_{Ki}^{Di} is the delay sensitivity of the gate to changes on its size, N is the number of paths passing through the gate and S_{Ki}^{Di} is the delay degradation of the gate.

Note that the metrics chosen for comparison are based on different characteristics of the gates. The metric in [16] focuses on identifying the gates suffering the largest degradation and affecting many paths. A similar metric has been proposed in [17] to improve the aged performance of critical paths in an ALU. The metric in [18] considers not only the gate delay degradation and the number of paths affected by the gate but also the delay sensitivity on gate sizing. This metric was shown to perform better than that proposed in [5]. The metrics in Eq. 12 were used in the proposed metric-guided design flow. Only the sizing-up heuristic was applied since the approaches in [16], and [18] do not consider a metric for sizing-down gates.

Table 5 shows the results for 20% and 10% of guardband constraint. For comparison purposes, the area overhead of our proposed approach is also given. It can be observed that our proposal gives designs with lower area overhead than those obtained using the metrics of [16] and [18]. This is because the proposed metrics includes important parameters not taken into account in the others such as the area impact and the paths slack. Furthermore, the proposed metric uses a statistical sensitivity that takes into

Table 4 Percentage of SP groups for which 10% of guardband may be violated

Circuit	Multiple WLs	Single WL
c880	4.79	55.11
c1908	0.11	32.24
c2670	3.52	53.62
c5315	4.17	29.7
c7552	0.05	24.77
s298	3.21	18.08
s838	0.33	9.71
s1423	0.29	0.48
s5378	4.57	51.98
Avg.	2.33	30.63

Table 5 Percentage of area overhead for target guardbands of 10% and 20% of using three selection and sizing methods

Circuit	Sizing: $GB_t = 20\%$			Sizing: $GB_t = 10\%$		
	Our	[16]	[18]	Our	[16]	[18]
c880	15.15	21.85	21.52	52.03	102.15	67.45
c1908	9.76	17.84	14.16	35.97	58.93	54.61
c2670	4.09	10.63	6.31	16.01	173.9	25.47
c5315	2.38	5.07	3.21	8.94	16.13	14.69
c7552	0.33	1.38	0.46	1.22	3.36	1.86
s298	6.26	10.61	11.21	16.34	40.01	24.30
s838	4.92	8.33	6.90	11.01	15.90	16.42
s1423	1.25	4.12	1.03	4.92	10.91	5.54
s5378	1.14	2.67	2.22	4.72	11.50	8.66
Avg.	5.03	9.16	7.44	16.79	48.08	35.44

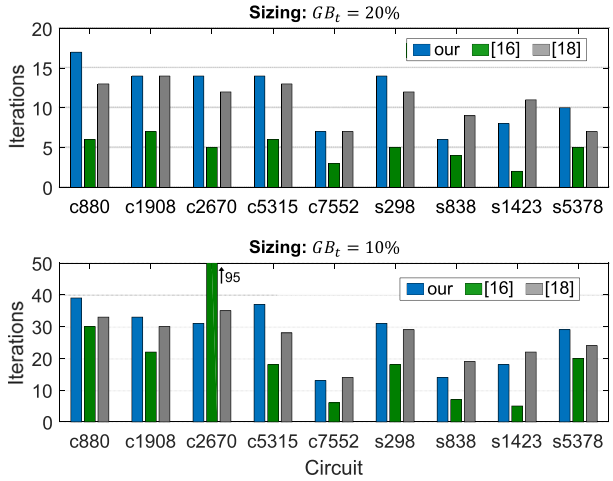


Fig. 10 Number of iterations to achieve target guardband

account the impact of sizing a gate on the nominal delay, delay deterioration and variability due to process variations. Among the other metrics, the one in [16] is less efficient for gate sizing. This is because this metric only considers the delay degradation and the number of paths impacted by the gate. However, it does not consider the path delay sensitivity to gate sizing. Therefore, it does not measure the potential delay improvement of sizing a gate. Although the metric in [18] includes the delay sensitivity parameters, this sensitivity does not consider aging or process variations effects. Therefore, it may fail to indicate the gates more beneficial for delay improvement.

Figure 10 shows the number of iterations performed when using each of the metrics for gate sizing. An iteration corresponds to the process of performing SSTA over all the PCPs to determine the current guarband required for the circuit and the Slow- and Fast- PCP subsets, the evaluation of the sizing metric for each gate in the PCPs, and the application of the sizing heuristic. It can be observed that the proposed metrics imply a larger number of iterations. This is because the proposed metrics select the gates giving an efficient delay-area trade-off, which are not necesarily the ones improving quicker the circuit delay. On the other hand, the metric in [16] gives a higher priority to those gates in the longest PCP of the circuit, which results in a quick delay reduction but with increased area overhead.

6 Conclusion

A gate sizing optimization methodology for guardband reduction in the presence of aging due to BTI and Process Variations have been presented in this paper. Since the workload that a circuit experiences over the lifetime if unknown at the design phase, the proposed methodology calculates the maximum realistic aged delay of the circuit paths for various workload profiles at main inputs, which define the stress probability of the devices. In such a way, the traditional worst BTI assumption and unreliable specific workload assumption have been avoided. It has been shown that a reasonable number of signal probability profiles is sufficient to obtain a good estimation of the maximum degraded delay of the circuit paths. For delay optimization towards the desired target guardband, gate metrics and a sizing heuristic have been proposed to select the best gates for both sizing-up to improve delay and sizing-down to mitigate area overhead. An approximation for the statistical sensitivity of a path delay has been proposed to mitigate computational effort of statistical timing analysis and speed-up metrics evaluation. The application of the proposed methodology on ISCAS benchmark circuits has shown that gate sizing using the proposed approach to estimate the maximum aged delay of the circuit paths results in significant area savings compared to gate sizing under worst BTI assumptions. Furthermore, it has been shown that the obtained designs can operate reliably for a different workload profile than those used during design optimization. The results using the proposed metrics has been compared against the results using other gates metrics in the literature, and it has been shown that the proposed approach provides a better area-delay trade-off.

Acknowledgments This work was supported by *CONACYT* (Mexico) through the Ph.D. scholarship number 420129/264560.

References

1. Kaczer B, Grasser T, Franco J, Luque MT, Weckx P, Roussel PJ, Groeseneken G (2012) Assessing reliability of nano-scaled cmos technologies one defect at a time. In: Proceedings International conference on emerging electronics, pp 1–2
2. Eghbalkhah B, Kamal M, Afzali-Kusha H, Afzali-Kusha A, Ghaznavi-Ghoushchi MB, Pedram M (2015) Workload and temperature dependent evaluation of bti-induced lifetime degradation in digital circuits. Microelectron Reliab 55(8):1152–1162
3. Bian S, Shintani M, Morita S, Awano H, Hiromoto M, Sato T (2016) Workload-aware worst path analysis of processor-scale nbti degradation. In: 2016 International great lakes symposium on VLSI (GLSVLSI), pp 203–208
4. Khan S, Hamdioui S, Kukner H, Raghavan P, Catthoor F (2012) Incorporating parameter variations in bti impact on nano-scale logical gates analysis. In: Proceedings IEEE international symposium on defect and fault tolerance in VLSI and nanotechnology systems (DFT), pp 158–163
5. Lu Y, Shang L, Zhou H, Zhu H, Yang F, Zeng X (2009) Statistical reliability analysis under process variation and aging effects. In: Proceedings design automation conference, 2009. DAC '09. 46th ACM/IEEE, pp 514–519
6. Alam MA, Roy K, Augustine C (2011) Reliability- and process-variation aware design of integrated circuits — a broader

perspective. In: Proceedings international reliability physics symposium, pp 4a.1.1–4a.1.11
7. van Santen VM, Amrouch H, Martin-Martinez J, Nafria M, Henkel J (2016) Designing guardbands for instantaneous aging effects. In: Proceedings of the 53rd annual design automation conference, DAC '16, pp 69:1–69:6, New York, NY, USA. ACM
8. Wu KC, Marculescu D (2009) Joint logic restructuring and pin reordering against nbti-induced performance degradation. In: Proceedings design, automation test in Europe conference exhibition, pp 75–80
9. Kiamehr S, Firouzi F, Tahoori MB (2012) Input and transistor reordering for nbti and hci reduction in complex cmos gates. In: Proceedings of the Great Lakes Symposium on VLSI, GLSVLSI '12, pp 201–206, New York, NY, USA. ACM
10. Abbas Z, Olivieri M, Khalid U, Ripp A, Pronath M (2015) Optimal nbti degradation and pvt variation resistant device sizing in a full adder cell. In: Proceedings international conference on reliability, infocom technologies and optimization (ICRITO) (trends and future directions), pp 1–6
11. Kiamehr S, Firouzi F, Ebrahimi M, Tahoori MB (2014) Aging-aware standard cell library design. In: Proceedings design, automation test in Europe conference exhibition (DATE), pp 1–4
12. Yabuuchi M, Kobayashi K (2016) Size optimization technique for logic circuits that considers bti and process variations. IPSJ Trans Syst LSI Design Methodol 9:72–78
13. Lin I-C, Syu S-M, Ho T-Y (2014) Nbti tolerance and leakage reduction using gate sizing. J Emerg Technol Comput Syst 11(1):4:1–4:12
14. Yang X, Saluja K (2007) Combating nbti degradation via gate sizing. In: Proceedings international symposium on quality electronic design (ISQED'07), pp 47–52
15. Khan S, Hamdioui S (2011) Modeling and mitigating nbti in nanoscale circuits. In: Proceedings IEEE 17th international on-line testing symposium, pp 1–6
16. Yang S, Wang W, Hagan M, Zhang W, Gupta P, Cao Y (2013) Nbti-aware circuit node criticality computation. J Emerg Technol Comput Syst 9(3):23:1–23:19
17. Kostin S, Raik J, Ubar R, Jenihhin M, Vargas F, Poehls LMB, Copetti TS (2014) Hierarchical identification of nbti-critical gates in nanoscale logic. In: Proceedings Latin American test workshop - LATW, pp 1–6
18. Jin S, Han Y, Li H, Li X (2011) Statistical lifetime reliability optimization considering joint effect of process variation and aging. Integration, the {VLSI}, J 44(3):185–191
19. Duan S, Halak B, Zwolinski M (2017) An ageing-aware digital synthesis approach. In: Proceedings 2017 14th international conference on synthesis, modeling, analysis and simulation methods and applications to circuit design (SMACD), pp 1–4
20. Gomez AF, Gomez R, Champac V (2018) A metric-guided gate-sizing methodology for aging guardband reduction. In: 2018 IEEE 19Th Latin-American test symposium (LATS), pp 1–6
21. Zafar S, Kumar A, Gusev E, Cartier E (2005) Threshold voltage instabilities in high- kappa; gate dielectric stacks. IEEE Trans Device Mater Reliab 5(1):45–64
22. Islam AE, Goel N, Mahapatra S, Alam MA (2016) Reaction-diffusion model, pp 181–207. Springer India, New Delhi
23. Sutaria KB, Velamala JB, Ramkumar A, Cao Y (2015) Compact modeling of BTI for circuit reliability analysis, pp 93–119. Springer, New York, p 1
24. Tudor B, Wang J, Chen Z, Tan R, Liu W, Lee F (2012) An accurate mosfet aging model for 28nm integrated circuit simulation. Microelectron Reliab 52(8):1565–1570. ICMAT 2011 - Reliability and variability of semiconductor devices and ICs
25. Yang HI, Yang SC, Hwang W, Chuang CT (2011) Impacts of nbti/pbti on timing control circuits and degradation tolerant design in nanoscale cmos sram. IEEE Trans Circuits Syst I: Regular Papers 58(6):1239–1251
26. Krishnappa SK, Singh H, Mahmoodi H (2010) Incorporating effects of process, voltage and temperature variation in bti model for circuit design
27. Jin S, Han Y, Li H, Li X (2010) *p*2 clraf An pre- and post-silicon cooperated circuit lifetime reliability analysis framework. In: Proceedings 19th IEEE asian test symposium, pp 117–120
28. Wang W, Reddy V, Bo Yang, Balakrishnan V, Krishnan S, Cao Y (2008) Statistical prediction of circuit aging under process variations. In: Proceedings 2008 IEEE custom integrated circuits conference, pp 13–16
29. Blaauw D, Chopra K, Srivastava A, Scheffer L (2008) Statistical timing analysis: from basic principles to state of the art. IEEE Trans Comput Aided Des Integr Circuits Syst 27(4):589–607
30. Xiong J, Zolotov V, He L (2007) Robust extraction of spatial correlation. IEEE Trans Comput Aided Des Integr Circuits Syst 26(4):619–631
31. Sivadasan A, Cacho F, Benhassain SA, Huard V, Anghel L (2016) Study of workload impact on bti hci induced aging of digital circuits. In: Proceedings 2016 design, automation test in europe conference exhibition (DATE), pp 1020–1021
32. White Paper Freescale (2008) Thermal analysis of semiconductor systems
33. https://www.synopsys.com

Andres Gomez received the electronics engineering degree (2011) from the Industrial University of Santander (UIS), Colombia, and obtained the M.Sc. (2013) and Ph.D. (2017) degrees in electronics sciences from the National Institute for Astrophysics, Optics and Electronics (INAOE), Mexico. He is a full-time professor at the Universidad Manuela Beltrán (UMB), Colombia. His current research interests include design of integrated circuits robust to reliability and process variations issues, test of integrated circuits for current and emerging technologies, and integrated circuit design for biomedical applications.

Victor Champac received the Ph.D. from the Polytechnic University of Catalonia (UPC), Spain. Since 1993 he is with the National Institute for Astrophysics, Optics and Electronics (INAOE-Mexico) where he is Titular Professor. Dr. Champac is IEEE Senior Member. He was co-founder of the Test Technology Technical Council-Latin America of IEEE Computer Society. He was the co-General Chair of the 2nd, 9th, 14th and 16th IEEE Latin American Test Workshop (symposium since 16th edition). He is member of the Board Director of Journal of Electronics Testing: Theory an Applications (JETTA). He participates in the Program Committee of several international conferences. He also serves as reviewer in several international conferences and journals. He has published over 120 papers at international conferences and journals. His research lines include: defect modeling in leading technologies, development of new test strategies for advanced technologies, aging reliable circuit design, and circuit design under process variations.