Efficient Transistor-Level Sizing Technique under Temporal Performance Degradation due to NBTI

Kunhyuk Kang, Haldun Kufluoglu, Muhammad Ashraful Alam and Kaushik Roy
Dept. of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA
{kang18, kufluogl, alam, kaushik}@ecn.purdue.edu

Abstract— Temporal performance degradation in VLSI circuits due to Negative Bias Temperature Instability (NBTI) has emerged as a challenging design issue in nano-scale technology. In this paper, we analyze the impact of NBTI degradation in circuit performance in terms of timing, and show that under worst case scenario, one can expect more than a 10% degradation in the maximum circuit delay after 3 years ($\sim 10^8$ seconds) operation time. Based on this observation, we propose an efficient transistor-level sizing algorithm based on a modified Lagrangian Relaxation (LR) technique to account for the temporal degradation of circuit and guarantee lifetime reliability of circuit under NBTI. The technique reformulates the sizing problem by considering the fact that only the rising ($0 \rightarrow 1$) delays of CMOS logic gates are affected by the NBTI. Experimental results on several ISCAS'85 benchmarks have shown that our proposed transistor-level sizing approach can reduce the area overhead of conventional cell-level sizing method by an average of 43%.

I. Introduction

For past decades, temporal reliability of MOSFET device has been considered as one of the key design factor in the device engineering field. Reliability degradation in MOSFET device can be due to several physical mechanisms such as Hot Carrier Injection (HCI), Negative Bias Temperature Instability (NBTI) [1], [2], Time Dependent Dielectric Breakdown (TDDB) and electromigration. Such degradations can cause performance degradations (e.g., timing or power) or even an unrecoverable malfunction in fabricated chips during its operation, requiring a well established reliability-aware device design techniques. However, as technology scaling moves towards the sub-100nm regime, sole device-level approaches may not be sufficient in achieving required level of reliability and hence there is need for co-effort from the circuit-level to improve reliability.

In this work, we focus on NBTI in PMOS transistors; one of the major dominant reliability degradation factors in nano-scale devices. In bulk MOSFET structure, undesirable Si dangling bonds can be generated due to crystal mismatch at the Si-SiO_2 interface, resulting in a generation of charged interfacial traps. Conventionally, to relax such mismatches, hydrogen passivation is applied to the Si surface before the oxidation process to transform dangling Si atoms to Si-H bonds. However, with time, these Si-H bonds can easily break during operation (i.e., on-state, negative gate bias for the PMOS) especially when nitrided oxides are used. Moreover, NBTI impact gets worse in scaled technology due to higher operation temperature and the usage of ultra thin oxide (i.e., higher oxide field). The broken bond generates interfacial traps and increases the threshold voltage (V_{Th}) of the device.

In [3], it was shown that the performance degradation in CMOS logic circuits due to NBTI degradation closely follows the trend of V_{Th} degradation in a single PMOS transistor. Further, they proposed a simple over-sizing method based on the Lagrangian Sizing (LR) [4] to compensate the degradation in maximum circuit delay and guarantee a lifetime functionality of the design. The method calibrates a worst-case degradation of V_{Th} in PMOS transistor in the initial design phase and computes an optimal sizing ratio for each and every cell (i.e., logic gate). However, considering the fact that the degradation due to NBTI only occurs in PMOS transistors, conventional cell-based sizing method may not be optimal in terms of total circuit area. In order to maximize the signal transfer efficiency (i.e., speed and power), CMOS logic gates are usually designed in such a way to balance the rising ($0 \rightarrow 1$) and falling ($1 \rightarrow 0$) delays. If PMOS V_{Th} increases (due to NBTI) in a CMOS logic gate, its rising delays are affected while its falling delays show only negligible difference. This means an additional timing slack can be found in the falling delays, and proper transistor sizing may achieve the required reliability with much smaller area overhead.

Based on this observation, in this paper, we propose an efficient transistor-level sizing algorithm under temporal NBTI degradations. Unlike conventional sizing algorithm, where a single sizing ratios is applied to each cell, we employ two different cell sizes for Pull-Up-Network (PUN) and Pull-Down-Network (PDN), respectively. The transistor-level sizing problem is then solved using a modified LR algorithm. Simulation results on several ISCAS'85 benchmark circuits show that by using our approach, we can reduce the area penalty (to compensate 3 years NBTI degradation) by an average of 43%, while retaining negligible changes in the design time.

The rest of the paper is organized as follows. In section 2, we explain the physics of NBTI and model the temporal V_{Th} degradation in PMOS transistors as a compact analytical form. Also, the impact of temporal V_{Th} degradation in circuit timing is discussed. In section 3, our reliability-aware transistor level sizing technique is proposed and explained in detail. Simulation results are presented in section 4. Finally, we conclude the paper in section 5.

II. Temporal Performance Degradation under NBTI

In this section, we analyze the impact of NBTI on temporal performance degradation at both device and circuit level. In the first part of this section, we setup an analytical expression for the temporal V_{Th} degradation in PMOS transistor due to NBTI based on the Reaction-Diffusion (R-D) framework proposed in [5]–[7]. Based on the transistor level degradation model, in the later part of the section, we will show how the impact of temporal V_{Th} increase is incorporated into the circuit level timing.

A. Temporal V_{Th} increase

NBTI is the result of trap generation at Si/SiO_2 interface in negatively biased PMOS transistors at elevated temperatures. The interaction of inversion layer holes with hydrogen-passivated Si atoms can break the $Si-H$ bonds, creating

1-4244-9707-X/06/$20.00 ©2006 IEEE

TABLE I

AC degradation factor α_S for different signal probabilities S_i. α_S scales down the bond breaking rate k_F.

Signal Probability (S_i)	AC degradation factor ($\alpha_S(S_i)$)
0.25	0.50
0.50	0.71
0.75	0.87

interface traps and H atoms, which can diffuse away from the interface (through the oxide) or can anneal an existing trap. An analytical model of interface trap generation has been modeled using the Reaction-Diffusion framework [2], [6], [7] and showed a power dependency on time with a fixed time exponent of 0.25. However, a sole H based model recently showed some discrepancy with the experimental measurements. Rather, it is now believed that the broken H atoms form H_2 molecules, which requires a new model.

General physical mechanism of H_2 based NBTI degradation is explained in [5], [7]. Generation of interfacial traps and the reverse annealing of Si-H bond can be expressed as follows,

$$\frac{dN_{IT}}{dt} = k_F(N_0 - N_{IT}) - k_R N_{IT} N_H^{(0)} \tag{1}$$

where N_{IT} is the density of interfacial trap, N_0 is the initial Si-H bond density and $N_H^{(0)}$ is the hydrogen density at the interface. k_F and k_R represent Si-H dissociation rate constant and reverse annealing rate, respectively. N_{IT} can be obtained by integrating the number of generated hydrogen molecules (H_2) inside the oxide and can be computed as,

$$N_{IT} = \int_0^{\sqrt{D_{H_2}t}} N_{H_2}(y,t)dy = \frac{N_{H_2}^{(0)}}{2}\sqrt{D_{H_2}t} \tag{2}$$

where t is the elapsed time, D_{H_2} is the diffusion coefficient of H_2. $N_{H_2}(y,t)$ and $N_{H_2}^{(0)}$ are the H_2 density in y (vertical depth toward the oxide) at time t and H_2 density at the interface ($y = 0$), respectively. Density of H_2 and H can be connected through the rate equation (i.e., $H + H \leftrightarrow H_2$) and can be expressed as,

$$k_1[H]^2 = k_2[H_2] \longrightarrow [H] = \sqrt{\frac{k_2}{k_1}[H_2]} \tag{3}$$

where k_1 and k_2 are the rate constants. Here, $[H]$ and $[H_2]$ are identical to the surface hydrogen density $N_H^{(0)}$ and molecule density $N_{H_2}^{(0)}$, respectively. Hence, by applying Eq. (2) to Eq. (3), surface hydrogen density can be expressed as,

$$N_H^{(0)} = \frac{\sqrt{2N_{IT}}}{(D_{H_2}t)^{1/4}}\left(\frac{k_2}{k_1}\right)^{1/2} \tag{4}$$

Finally, by merging Eqs. (1) and (4), we can solve for N_{IT} as,

$$N_{IT}(t) = \left(\frac{k_1}{2k_2}\right)^{1/3}\left(\frac{k_F N_0}{k_R}\right)^{2/3}(D_{H_2}t)^{1/6} \propto t^{1/6} \tag{5}$$

where we can observe that the trap generation has a power dependency on time with a fixed exponent of 1/6.

In a real circuit operation, the effective on-time of transistors are bounded by its input signal probability. In our work, we define the Signal Probability S_i at the input of gate i as a fraction of operating cycle which contributes to the NBTI degradation, that is, logic low in CMOS since PMOS transistors mainly get affected by NBTI. Depending on the S_i value, bond-breaking rate k_F is being scaled down by the AC degradation

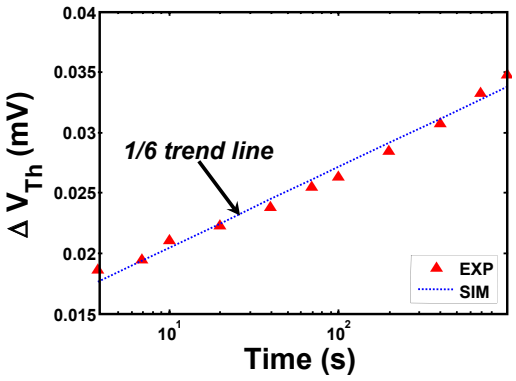

Fig. 1. Comparison between experimental data from [8] and our proposed model.

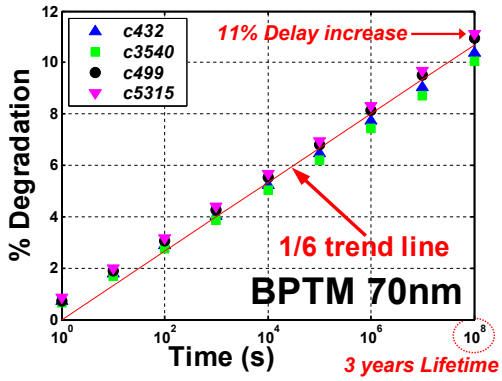

Fig. 2. Temporal delay degradation for 3 years in several ISCAS'85 benchmark circuits.

factor α_S. α_S values for various S_i's are computed using the R-D framework [7]. Table I shows several AC degradation factors for different signal probabilities. Considering this, trap generation N_{IT} can be now transformed into an increase in V_{Th} as follows,

$$\Delta V_{Th}(t) = (m+1)\frac{qN_{IT}(t)}{C_{OX}} = K_C \times \alpha_S(Si)^{2/3} \times t^{1/6} \tag{6}$$

where m is a mobility degradation factor and K_C is a constant factor from Eq. (5). Fig. 1 shows a comparison between our model and the experimental data from [8] where we can observe a good match over a wide range as well as the power dependency of V_{Th} degradation with a fixed time exponent of 1/6.

B. Performance degradation

Using the temporal V_{Th} model proposed in the previous section, we can now estimate the delay degradation in the circuit. It was shown in [3] that the increase in circuit delay also follows the same exponent of V_{Th} degradation. V_{Th} model introduced in Eq. (6) was integrated into the delay model and Static Timing Analysis (STA) was used to compute the worst-case maximum delay of a given circuit. All relevant parameters in Eq. (6) are calibrated for the BPTM 70nm technology node [9]. Fig. 2 shows the temporal delay degradations for several ISCAS benchmark circuits. As expected, the degradation in delay also shows a power dependency to time with a fixed exponent of 1/6. In a 3 year time period, we can observe up to 11% increase in delay. In a design with very tight timing

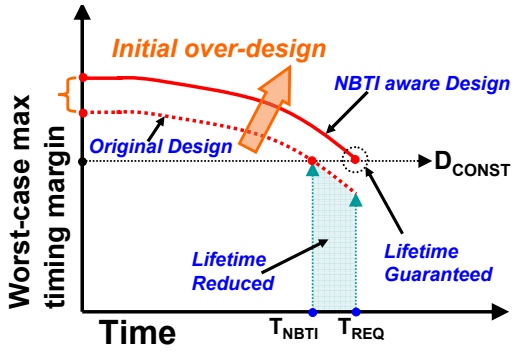

Fig. 3. NBTI-aware sizing method.

TABLE II
Rising delay degradation in several standard cells. Each cell is designed using BPTM 70nm technology file.

Logic Cell	fanin	Delay (ps)		Degradation (%)
		$t = 0$	3 years	
INV	1	13.77	16.77	21.81
NAND	2	16.86	19.88	17.93
NAND	3	19.57	22.45	14.75
NOR	2	17.26	21.89	26.79
NOR	3	23.80	30.19	26.87

margins, the increase in the critical delays can result in a timing failure. Based on the above observations, we propose an efficient sizing method to compensate the impact of NBTI and guarantee a lifetime functionality of the design.

III. Reliability Aware Transistor-Level Sizing

In this section, we propose our reliability-aware transistor-level sizing algorithm based on the Lagrangian Relaxation (LR) [4] technique. The basic idea of our reliability-aware sizing closely follows that proposed in [3] and is shown in Fig. 3. As can be seen, the timing margin of the design reduces with time, and after certain time (T_{NBTI}), the design fails to meet the constraint (D_{CONST}), resulting in a timing failure before its required lifetime T_{REF}. To ensure a lifetime reliability ($T_{NBTI} > T_{REQ}$), we first calibrate the estimated V_{Th} degradation in each transistor using the model introduced in section 2.1. Then, the circuit is optimally sized based on each transistor having pre-calibrated amount of V_{Th} increase.

As mentioned in the introduction, the original LR based sizing proposed in [4] applies a single sizing ratio for each logic gate (i.e., x_i for ith gate in the circuit). However, with the impact of NBTI, PUN side of the gates are imposed to more degradation than the PDN side, resulting in a large skew between the rising and the falling delays. Table II shows the degradation in rising delays of several standard cells. Results are obtained from HSPICE simulation using the BPTM 70nm files [9]. In contrast to a large reduction in the rising delays, falling delays only show negligible changes. Hence, if we target to size the circuit for the max delays (i.e., which would be the rising delay under NBTI), it can result in an extra timing slack for the falling delays. Hence, in order to efficiently cope with the different timing slacks for rising and falling delays, we applied different sizing factor for PUN and PDN, respectively. The details of our method is further explained in the following context.

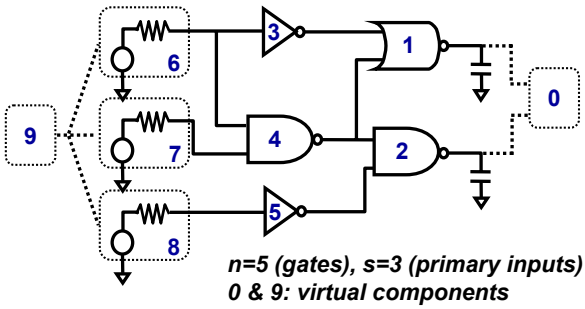

Fig. 4. Circuit representation for the LR based transistor-level sizing algorithm.

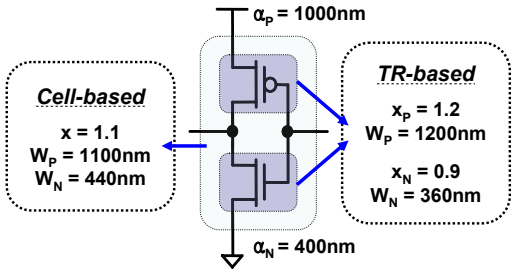

Fig. 5. Cell-based and TR-based sizing example.

A. Basic Notations

Basic notations used in our algorithm is introduced in this section. Fig. 4 represents an example circuit for our sizing algorithm. The circuit consists of n logic gates, and s primary inputs. In addition, there are two virtual component sourcing all primary inputs (component 9) and sinking all primary outputs (component 0). All components (total of $n + s + 2$) are numbered in its reverse logical order. Latest rising arrival time and falling arrival time at the output of the ith gate are r_i and f_i, respectively. Maximum rising delay and falling delays at ith gate are $D_{r,i}$ and $D_{f,i}$, respectively.

B. Transistor-level sizing using LR

Our transistor-level sizing algorithm employs two sizing factors; $x_{N,i}$ for PDN and $x_{P,i}$ for PUN of ith gate, respectively. An example of a simple inverter is shown in Fig. 5, where we can observe the difference of cell-based and transistor-based sizing. Under this assumption, the problem of minimizing total area subject to a delay constraint can be formulated as follows,

Sizing for Minimum Area

$$\begin{aligned} \text{Minimize} \quad & \textstyle\sum_{i=1}^{n}(\alpha_{N,i}x_{N,i} + \alpha_{P,i}x_{P,i}) \\ \text{Subject to} \quad & f_p \le A_0,\ r_p \le A_0 \qquad \forall p \in P \\ & L_{N,i} \le x_{N,i} \le U_{N,i},\ L_{P,i} \le x_{P,i} \le U_{P,i} \end{aligned}$$

where n is the number of logic gates in the circuit, $\alpha_{N,i}$ and $\alpha_{P,i}$ are the basic sizes of the PDN and the PUN parts (Fig. 5), respectively. A_0 is the maximum delay constraint and P is a set of all the primary output edges. $L_{N,i}$ and $U_{N,i}$ represent the minimum and maximum achievable size factor for the PDN of ith logic gate ($L_{P,i}$ and $U_{P,i}$ for the PUN). It was shown in [4] that the computational complexity of above problem has exponential dependence to the number of gates n. Hence, to relax the complexity, the delay constraint at the primary output (A_0) is transformed into delay constraint at each logic gate, noted as Primal Problem (PP) in [4]. In our work, we

$$L_\lambda(x,r,f) = \sum_{j\in in(0)} \Big(\lambda_{r,j0}(f_j - A_0) + \lambda_{f,j0}(r_j - A_0)\Big) + \sum_{i=n+1}^{n+s} \Big(\lambda_{r,mi}(D_{r,i} - r_i) + \lambda_{f,mi}(D_{f,i} - f_i)\Big)$$
$$+ \sum_{i=1}^{n} \sum_{j\in in(i)} \Big(\lambda_{r,ji}(f_j + D_{r,i} - r_i) + \lambda_{f,ji}(r_j + D_{f,i} - f_i)\Big) + \sum_{i=1}^{n}(\alpha_{N,i}x_{N,i} + \alpha_{P,i}x_{P,i}) \quad (7)$$

modified the PP into a modified Dual Primal Problem (DPP) as follows,

Dual Primal Problem (DPP)
Minimize $\sum_{i=1}^{n}(\alpha_{N,i}x_{N,i} + \alpha_{P,i}x_{P,i})$
Subject to
$f_j \le A_0,\ r_j \le A_0 \quad j \in \text{in}(0)$ /*outputs*/
$f_j + D_{r,i} \le r_i,\ r_j + D_{f,i} \le f_i \quad i = 1,...,n \quad \forall j \in \text{in(i)}$
$D_{r,i} \le r_i,\ D_{f,i} \le f_i \quad i = n+1,...,n+s$ /*inputs*/
$L_{N,i} \le x_{N,i} \le U_{N,i},\ L_{P,i} \le x_{P,i} \le U_{P,i} \quad i = 1,...,n$

Compared to the original form of PP, constraints are written for both the rising and falling arrival times in the DPP.

To solve the problem, DPP is first transformed into a polynomial equation shown in Eq. (7) using two sets of Lagrangian multipliers; λ_r and λ_f for rising and falling signals, respectively. $\lambda_{r,ji}$ represents the Lagrangian multiplier for the falling output arrival time of gate j driving gate i (correspondingly rising) and vice versa for the $\lambda_{f,ji}$. $\lambda_{f,mi}$ and $\lambda_{r,mi}$ are the multipliers for the virtual source node (i.e., $m = n + s + 2$). Now, DPP is identical to minimizing Eq. (7) under the cell level sizing constraint (i.e., $L_{N,i} \le x_{N,i} \le U_{N,i}$ and $L_{P,i} \le x_{P,i} \le U_{P,i}$). Eq. (7) can be further simplified if both λ_r, λ_f's follows the following condition,

Dual Optimality Condition (DOC): for $1 \le k \le n+s$
$$\sum_{i\in out(k)} \lambda_{r,ki} = \sum_{j\in in(k)} \lambda_{f,jk}, \quad \sum_{i\in out(k)} \lambda_{f,ki} = \sum_{j\in in(k)} \lambda_{r,jk}$$

In contrast to the original Optimality Condition shown in [4], DOC connects a relationship between the input and output timing signals while considering the type of transition (e.g., rising output to falling input and vice versa). By ensuring the DOC, Eq. (7) can be rewritten by employing the Kuhn-Tucker condition as follows,

$$L_\lambda(x,r,f) = \sum_{i=1}^{n+s} \Big(\mu_{f,i}D_{f,i} + \mu_{r,i}D_{r,i}\Big) - (\mu_{r,0} + \mu_{f,0})A_0$$
$$+ \sum_{i=1}^{n}(\alpha_{N,i}x_{N,i} + \alpha_{P,i}x_{P,i}) \quad (8)$$

where $\mu_{r,i}$ is the sum of λ_f's at all inputs of ith gate (i.e., $\mu_{r,i} = \sum_{j\in in(i)} \lambda_{f,ji}$). Similarly, $\mu_{f,i}$ can be obtained by the sum of λ_r's from the input.

Given Eq. (8), size ratios for the minimum $L_\lambda(x,r,f)$ can be obtained by greedy local resizing method [4], [10]. In our work, we applied Sakurai's model [11] to compute the gate delay over each input pin to output pin combination. Using the Sakurai's delay model, rising and falling delays over gate i can be represented as follows,

$$D_f = \frac{\gamma_1}{x_{N,i}} + \gamma_2 x_{N,i} + \gamma_3 x_{P,i} \quad (9)$$
$$D_r = \frac{\gamma_4}{x_{P,i}} + \gamma_5 x_{P,i} + \gamma_6 x_{N,i}$$

where $\gamma_1 \sim \gamma_2$ represents the pre-characterized process dependent constants [11]. Conceptually, Eq. (9) shows how each sizing components are related to rising and falling delays at a single gate. For example, since both falling and rising delays can have linear dependencies with respect to the input capacitance [12], $x_{N,i}$ and $x_{P,i}$ are factored by a linear constants γ_2, γ_3, γ_5 and γ_6. By inserting Eq. (9) into Eq. (8), and differentiating with respect to $x_{N,i}$ (and to $x_{P,i}$). We can obtain the optimal size ratio for PDN and PUN to be,

$$x_{N,i} = min\Big(U_{N,i}, max\Big(L_{N,i}, \sqrt{\frac{\gamma_1\mu_{f,i}}{\gamma_2\mu_{f,i} + \gamma_6\mu_{r,i} + \alpha_{N,i}}}\Big)\Big) \quad (10)$$
$$x_{P,i} = min\Big(U_{P,i}, max\Big(L_{P,i}, \sqrt{\frac{\gamma_4\mu_{r,i}}{\gamma_5\mu_{r,i} + \gamma_3\mu_{f,i} + \alpha_{P,i}}}\Big)\Big)$$

The greedy sizing applies the optimal sizing ratios given by Eq. (10) iteratively until there is no improvement (i.e., no change in the size factor itself). Once the optimal size ratios for given λ's are obtained, the quality of current solution is measured by Eq. (8). If the $L_\lambda(x,r,f)$ is minimized below some error boundary, the circuit is optimized. Otherwise, we move the current λ's following the subgradient direction. Both λ_r's and λ_f's are moved by multiplying step factor ρ_k's to the subgradient direction and adding it back to λ. After the move, both λ_r's and λ_f's are project back to the nearest λ's satisfying the DOC in order to utilize Eq. (8). Algorithm 1 depicts our transistor-level sizing algorithm, noted as the Dual LR Sizing. ERR_λ and ERR_{Δ_x} represent the error boundary for closing the LR sizing and greedy local resizing steps, respectively.

Algorithm 1 Dual LR Sizing

```
Inputs: Delay constraint A_0, Error bounds ERR_λ and ERR_Δx
while λ(r,f,x) ≤ ERR_λ do
    % Greedy Local Resizing
    while Σ Δ_x ≤ ERR_Δx do
        for each logic gate in a reverse logical order do
            Apply optimal size for PUN and PDN using Eq. (10)
        end for
        Measure overall changes in size ratio Σ Δ_x
    end while
    % Move λ's
    Compute new set of λ_r's and λ_f's while satisfying the Dual Optimality Condition
end while
return optimal PUN and PDN size x_N and x_P for all gates in the circuit
```

The runtime complexity of our algorithm has linear dependency with the number of gates in the circuit. In addition, the optimality and convergence of our algorithm can be verified in a similar manner as shown in [4]. However, due to the limited space, detailed derivation and proof are not included in this paper.

C. NBTI-aware sizing method

Using the transistor-level sizing framework proposed in the previous section, we now introduce our NBTI-aware sizing

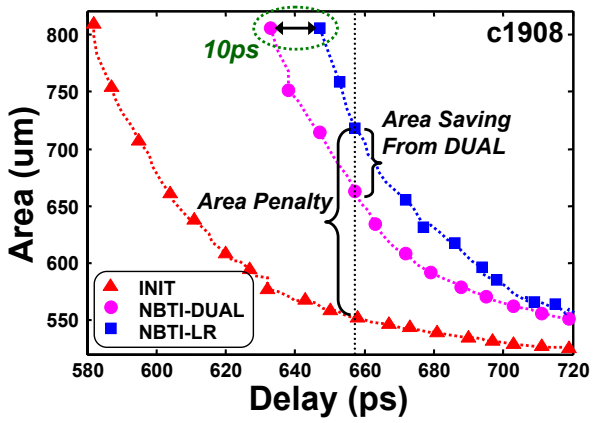

Fig. 6. Delay versus area curve of c1908 from 1) INIT: LR sized initial ($t = 0$) design, 2) NBTI-LR: LR sized for 3 years lifetime and 3) NBTI-DUAL: DUAL sized for 3 years lifetime.

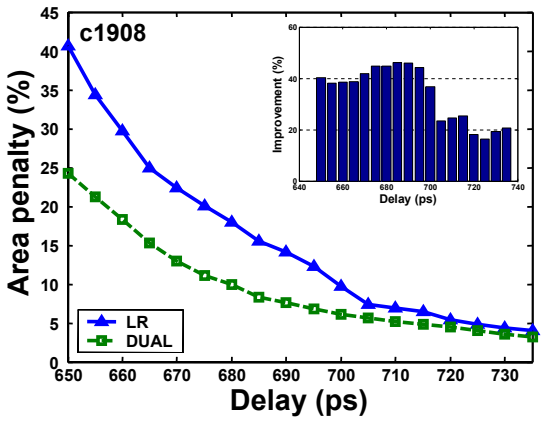

Fig. 7. Area penalty using LR and DUAL sizing in c1908 for different delay constraints. Corresponding improvements using DUAL algorithm (small figure).

method. As mentioned earlier, the basic idea of our method is to pre-calibrate the V_{Th} degradation in each PMOS transistor due to the NBTI considering the signal probability at the gate input. Then, sizing is applied while assuming that the circuit has degraded over the lifetime (T_{REF}) period (i.e., refer Fig. 3). The following flow explains the overall design methodology.

1) Setup the two design constraints: Delay Constraint A_0, Required Lifetime T_{REQ}.
2) Calibrate the signal probability (S_i, fraction of on-time for PMOS transistors) at each gate output.
3) Compute V_{Th} degradation for T_{REQ} time using Eq. (6) in each node considering S_i. Replace the nominal V_{Th} with the degraded one.
4) Apply Dual LR Sizing (Algorithm 1).
5) Obtain a NBTI-aware optimal design

In reality, after applying the sizing at the 4th step, timing information at each node can change from its initial values. Correspondingly, signal probability at those nodes can be different from the value extracted at step 2. Hence, an iterative loop between step 2 and 4 should be applied to converge the solution to a point where the sized value at step 4 only alters the pre-computed signal probability within a minimum range. However, in [3], it was shown that the sensitivity of NBTI degradation with respect to the signal probability is small, thus in our work, we neglect the change in signal probability after the sizing.

IV. Experimental Results

The proposed sizing algorithm was implemented in C, noted as the DUAL sizing. A set of ISCAS'85 benchmark circuits were chosen for the simulation. All the circuits were synthesized using the LEDA standard cell library and properly scaled down for the BPTM 70nm [9] technology node.

Fig. 6 depicts 3 different delay versus area curves obtained for circuit c1908. Delay was measured using Static Timing Analysis (STA), while areas were extracted as a sum of channel width from all the gates in the circuit. First, the curve INIT depicts the sizing (LR) result when NBTI is not considered (i.e., identical to the result at $t = 0$). NBTI-LR and NBTI-DUAL shows a sizing result considering 3 years degradation from NBTI when the original LR algorithm and our proposed DUAL sizing algorithm are applied, respectively. For a delay constraint of 655ps, the conventional LR sizing based compensation requires approximately 30% overhead compared to INIT. However, using our proposed DUAL sizing method, we can reduce the amount of penalty by nearly half in this case. Fig. 7 depicts the area penalty using LR and DUAL sizing for different delay constraints. In general, it can be observed that the improvements of DUAL sizing becomes more prominent when delay constraints are more stringent. Also, DUAL sizing reduces the minimum achievable delay by 10ps. Note that, all these improvements are stemming from utilizing the additional timing slack in falling delays by our transistor-level DUAL sizing method. For the c1908 example shown above, we measured in average a 17% skew in the PUN and PDN size ratios.

Table III summarizes our simulation results. All benchmark circuits were initially (i.e., for $t = 0$) sized using LR (INIT). The resulting size value is shown in the third column. Then, for each circuit, 3 year degradation in PMOS transistors due to the NBTI is applied. For the degraded circuits, we applied both the original LR algorithm and our proposed DUAL sizing algorithm (Section 4). As mentioned in section 4, the effectiveness of our proposed technique can vary over various delay constraints. Hence, the sizing was applied to three different corner of delay constraint. FAST corner (column 4∼6) represents the most stringent constraint where the slope of tangent to the area versus delay curve was more than 5. Similarly, NORM (column 7∼9) and SLOW (column 10∼12) constraints were determined at the point where the slope of tangent was 1 and 0.2, respectively. On average, area saving of 54, 45 and 31% were achieved for FAST, NORM and SLOW corner, respectively, using our algorithm. As predicted, the effectiveness of our algorithm increases when the delay targets are stringent.

Note that the effectiveness of the DUAL sizing algorithm also applies to any general cases where NBTI is not considered (i.e., since it is practically impossible to equalize rising and falling delays at all timing nodes). However, the benefits from using DUAL sizing is the most prominent when NBTI is considered. Fig. 8 depicts the % area penalty of LR and DUAL sizing for different degradation time in c2670 circuit. As the degradation time reduces (and NBTI is less), difference between rising and falling delays reduces, and as a result, effectiveness of DUAL sizing reduces.

Also, runtime of DUAL sizing algorithm was measured and compared to that of the original LR algorithm. On average, DUAL sizing was able to converge to an optimal solution with less than 5% increase in the runtime.

TABLE III
Simulation results for ISCAS'85 benchmark circuits

Circuit	No. of Gate	Init. Area (um)	Area overhead (%)								
			FAST			NORM			SLOW		
			LR	DUAL	saving	LR	DUAL	saving	LR	DUAL	saving
c74182	23	47.46	19.43	6.78	65.10	4.37	2.85	34.76	1.85	1.26	31.66
c74L85	38	66.36	29.69	17.50	41.03	10.27	4.12	59.90	2.25	1.22	45.73
c74283	50	91.56	68.76	31.05	54.84	44.20	21.93	50.39	39.70	17.40	56.18
c74181	95	173.88	38.72	7.47	80.70	11.44	2.83	75.27	2.21	1.15	47.96
c432	142	295.05	31.46	12.53	60.16	4.93	3.40	30.92	1.88	1.81	4.00
c1908	452	731.85	56.75	29.18	48.58	14.95	7.87	47.38	5.01	3.83	23.58
c880	511	746.97	14.37	8.33	42.09	4.27	3.10	27.50	1.54	1.08	30.08
c499	516	797.58	42.91	35.94	16.23	18.33	12.57	31.45	8.26	5.75	30.42
c2670	841	1115.10	21.69	8.28	61.83	4.64	1.71	63.15	0.97	0.59	39.81
c3540	932	1796.97	4.86	2.99	38.48	1.59	1.17	26.23	0.56	0.53	5.12
c5315	1900	2812.74	18.37	5.26	71.36	2.24	1.10	50.69	0.42	0.27	36.69
c6288	2421	4171.02	29.32	10.06	65.69	19.17	10.91	43.09	3.02	2.23	26.16
Average			31.36	14.61	53.84	11.70	6.13	45.06	5.64	3.09	31.45

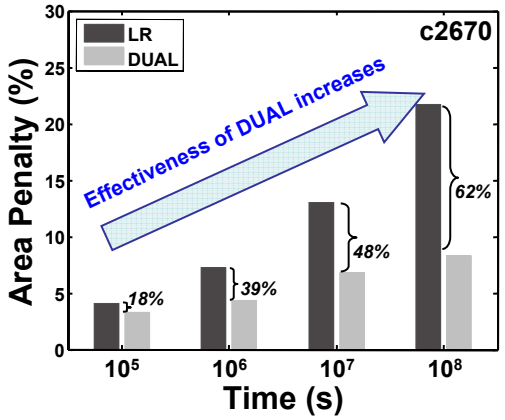

Fig. 8. % Area penalty of LR and DUAL sizing for various degradation time.

V. Conclusion

Time-dependent reliability degradation due to the Negative Bias Temperature Instability (NBTI) can have severe impacts on the circuit performance over time. In this paper, we showed that the maximum circuit delay can increase more than 10% over a 3 year time period due to the NBTI. To compensate the impact of NBTI on the performance, we proposed an efficient transistor-level sizing algorithm, referred as the DUAL-sizing technique. The algorithm utilizes the fact that under NBTI, only the PMOS transistors are affected, resulting in a large skew between the rising and falling delays in CMOS logic gates. Hence, different sizing ratios are applied to the PDN and PUN side of the gates. The problem of computing the optimal PDN and PUN size ratios are solved using our proposed modified DUAL Lagrangian Relaxation (LR) techniques. Simulation results on several ISCAS'85 benchmark circuits showed that compared to the conventional cell-based LR sizing, the proposed method can reduce the area overhead of compensating the 3 year NBTI by an average of 43%.

References

[1] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," Journal of Applied Physics, vol. 94, no. 1, pp. 1–8, july 2003.

[2] K.O. Jeppson and C.M. Svensson, "Negative bias of MOS devices at high electric fields and degradation of MNOS devices," Journal of Applied Physics, vol. 48, pp. 2004–2014, 1977.

[3] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Impact of NBTI on the temporal performance degration of digital circuits," IEEE Electron Device Letter, vol. 26, no. 8, pp. 560–562, august 2005.

[4] C. P. Chen, C. C. N. Chu, and D. F. Wong, "Fast and exact simultaneous gate and wire sizing by Lagrangian relaxation," IEEE Transactions on Computer-Aided Design of Integrated Circuits and System, vol. 18, no. 7, pp. 1014–1025, 1999.

[5] H. Kufluoglu and M. A. Alam, "Theory of interface-trap-induced NBTI degradation for reduced cross section MOSFETs," IEEE Transactions on Electron Devices, vol. 53, no. 5, pp. 1120–1130, may 2006.

[6] M. A. Alam, "A critical examination of the mechanics of dynamic NBTI for PMOSFETs," in International Electron Device Meeting, 2003, pp. 346–349.

[7] S. Chakravarthi, A. Krishnan, V. Reddy, C. F. Machala, and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," in IEEE International Reliability Physics Symposium Proceedings, 2004, pp. 273–282.

[8] A. T. Krishnan, C. Chancellor, S. Chakravarthi, P. E. Nicollian, V. Reddy, A. Varghese, R. B. Khamankar, and S. Krishnan, "Material dependence of hydrogen diffusion: implications for NBTI degradation," in IEEE International Electron Device Meeting, Tehcnical Digest, 2005, pp. 688–691.

[9] Berkeley, Predictive Technology Model, http://www-device.eecs.berkeley.edu/∼ptm, 1996.

[10] C. P. Chen, H. Zhou, and D. F. Wong, "Optimal nonuniform wire-sizing under the elmore delay model," in IEEE/ACM International Conference on Computer Aided Design, 1996, pp. 38–43.

[11] T. Sakurai and R. Newton, "Delay analysis of series-connected mosfet circuits," IEEE Journal of Solid-State Circuits, pp. 122–131, 1991.

[12] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, Digital Integrated Circuits, Prentice Hall, 2nd edition, 2002.