NBTI Aware Voltage Scaling Using Linear Transient Weighted Factors

Jaeil Lim, Hyejeong Hong, Sungho Kang
School of Electrical & Electronic Engineering, Yonsei University, Seoul, Korea
limji, hjhong@soc.yonsei.ac.kr, shkang@yonsei.ac.kr

Abstract

Negative Bias Temperature Instability (NBTI) effect is considered as a main reliability constraint on Very Large Scaled Integration (VLSI) design. NBTI aware voltage scaling has been researched to mitigate the NBTI effect. In this work, a new voltage scaling technique is proposed which selects a proper voltage level considering both stress intensity and immediate power consumption whereas early developed techniques only consider immediate power consumption. The proposed method suppresses the stress intensity at the beginning of lifetime and elevates the power efficiency over the lifetime. This method takes benefit on overall power consumption, roughly 12.9% reduction on uniform stress compared with the existing techniques.

Keywords: VLSI, reliability, NBTI, voltage scaling.

1. Introduction

Technology scaling induces many changes such as increased IC temperature and lower voltage. Accordingly, VLSI reliability issues become very important. NBTI is becoming a main reliability issue because of very sensitive acceleration on high temperature, high electric field in the gate oxide and reduced difference between V_{dd} and V_{th}. The NBTI degradation will be a more important reliability issue in the near future because there is a new developing integration technology that is 3-Dimentional Integrated Circuit (3D IC). This technique raise the heat generation drastically as the system performance is greatly improved.

To mitigate the NBTI degradation, voltage scaling techniques have been researched which monitor and compensate the degradation. Tiwari et al. [1] proposed a basic mechanism of NBTI aware voltage scaling. Karpuzcu et al. [2] proposed a reliable many-core system managing technique with voltage scaling. Zhang and Dick [3] proposed a method that discrete voltage levels can be applied on simulation method. All of these techniques scale the voltage using supply voltage (V_{dd}) only. There are some voltage scaling techniques that use both V_{dd} and body bias voltage (V_{bs}) [4], [5]. These techniques utilize V_{bs} for voltage scaling since the enhanced V_{dd} worsens the NBTI stress intensity, but enhancing V_{bs} becomes less serious. Mintarno et al. [6] proposed a pre-computed optimal voltage selection method. This technique assumes that NBTI stress intensity is predictable and uniform. However, in most cases, system workload and board temperature are not uniform during lifetime. In addition, NBTI stress intensity fluctuates. Generally, the NBTI degradation is sensed or calculated and reflected to control systems in real time. However, this pre-computed optimal voltage selection method cannot reflect the degradation data, because the worst case stress is assumed and voltage selection cannot adjustable on real time. The difference between the worst case NBTI degradation and the real degradation will cause excessive power consumption. In this work, a new technique is proposed which shows almost ideal performance. In addition, real time stress intensity fluctuation can be reflected to voltage selection. Thus this method incurs no excessive power consumption.

2. NBTI degradation model

The NBTI effect can be accounted with Reaction Diffusion (RD) model. The NBTI degradation progresses when PMOS have negative bias input on gate to source voltage. And it is accelerated by high temperature sensitively. Every PMOS in a logic circuit repeats stress and recovery sequence depending on the gate input signal change. Fluctuant increasing amount of threshold voltage is arranged as upper bound model as [7], [8]:

$$\Delta V_{th} \le \left(\frac{\sqrt{K_v^2 \alpha T_{clk} / min(\alpha, 1-\alpha)}}{1-\beta_m^{1/2n}} \right)^{2n} \quad (1)$$

$$\beta_m = 1 - \frac{2\varepsilon_1 t_e + \sqrt{\varepsilon_2 C (1-\alpha) T_{clk}}}{2t_{ox} + \sqrt{Ct}} \quad (2)$$

$$C = exp(-E_a/kT)/T_0 \quad (3)$$

$$K_v = \left(\frac{q t_{ox}}{\epsilon_{ox}}\right)^3 {K_1}^2 C_{ox} (V_{gs} - V_{th}) \sqrt{C} exp\left(\frac{2E_{ox}}{E_{o1}}\right) \quad (4)$$

where α is ratio of stress to stress-recovery cycle, T_{clk} is the time period of one stress-recovery cycle, T is

temperature, V_{gs} is gate to source voltage on stress phase and V_{th} is threshold voltage. Refer [7], [8] for more detailed information.

Increase of circuit delay is modeled as [9]. Voltage scaling technique increases the supply voltage to reduce the circuit delay and compensate the NBTI degradation. The circuit delay is modeled as:

$$\tau \propto \frac{C_L V_{DD}}{(V_{DD}-V_{th})^{\alpha}} \tag{5}$$

$$V_{th} = V_{th0} - K_1 V_{DD} - K_2 V_{bs} \tag{6}$$

where τ is the circuit delay, C_L is the load capacitance, V_{DD} is the supply voltage, V_{th} is the threshold voltage and α is a velocity saturation index, V_{th0} is reference threshold voltage, K_1 and K_2 are technology constant from [9].

The circuit temperature can be modeled by thermal resistance [3]. For NBTI aware voltage scaling technique, voltage control period is sufficient long (scale of days). It's sufficient time for temperature saturation. So, in this work, dynamic thermal change is not considered. Temperature and power consumption [9] can be modeled as:

$$T = T_a + P\,R_t \tag{7}$$

$$P_{\mathrm{ac}} = C_{eff} V_{dd}^2 f \tag{8}$$

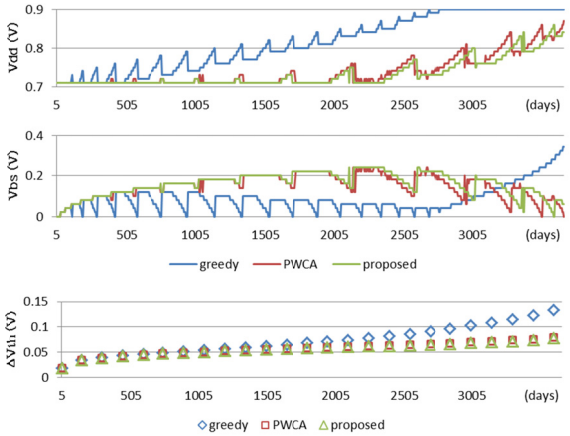
$$P_{dc} = V_{dd} I_{sub} + |V_{bs}| I_j \tag{9}$$

$$I_{sub} = K_3 \mathrm{e}^{K_4 V_{dd}} \mathrm{e}^{K_5 V_{bs}} \tag{10}$$

where T_a is the ambient temperature, P is a power consumption which defined as $P = P_{ac} + P_{dc}$, R_t is a thermal resistance, P_{ac} is dynamic power, C_{eff} is effective switching capacitance, f is system frequency, P_{dc} is static power, I_{sub} is subthreshold leakage, I_j is bias junction current and K_3 through K_5 are technology constant from [9].

3. Proposed method

The voltage scaling technique compensates the NBTI degradation with voltages and that induces the extra power consumption. If NBTI degradation progresses at the beginning of lifetime, more power resources should be consumed to compensate the precedent degradation during the rest of the lifetime. On the other hand, at latter part of the target lifetime, current power consumption is significant, because there is a few following terms. In this work, a new voltage scaling technique is proposed which trades the stress intensity and the immediate power consumption depending on current time progress versus target lifetime.



Stress intensity and the immediate power consumption can be traded by controlling various scaling factors like V_{dd} and V_{bs}. When V_{bs} is scaled up, V_{th} and circuit delay are reduced and the NBTI degradation is compensated. The V_{bs} control incurs a leakage power increment, so V_{bs} should be scaled up carefully. Scaling V_{dd} up incurs a less power increment but a stronger NBTI stress, compared with V_{bs} scaling up. We have performed simulations with these two factors. Anything schedulable can be combined with this method, such as cooling power.

Figure 1 shows a NBTI aware voltage scaling example during ten years (3650 days) with applying the maximum stress intensity uniformly. The goal of the example is to minimize the power consumption over the ten years with the guarantee of ten-year normal operations. We assumed that voltage selection and transition is performed at every five-day cycle, and both V_{dd} and V_{bs} have twenty discrete voltage levels. PWCA (Progressive-Worst-Case-Aging) [6] is an overall power optimization method with the prediction of a uniform stress. Greedy [5], [6] is an early developed technique that selects voltages which is predicted to minimize power during upcoming five days.

The proposed method and the greedy select the voltage levels with prediction of upcoming five days, whereas PWCA predicts whole lifetime. Computation of the following lifetime prediction is very complex, because there is extremely large number of cases of voltage transition. So the voltage scaling system cannot compute PWCA every five-day cycle. In other words, PWCA selects voltage levels in design time, so the data which are monitored in real time cannot be reflected. That means PWCA cannot recognize real stress intensity which is generally smaller than the maximum stress intensity. PWCA fixes the voltage selection at design time based on the maximum stress intensity. Thus, PWCA selects excessive high voltages for degradation state and brings extra power consumption. On the other hand, the computation of upcoming five days prediction is reasonable to be computed in real time. Therefore, unlike

Table 1: Comparison of results

Method	greedy	PWCA	proposed
Power(normalized)	1	0.870008	0.870526
Data reflection	Possible	Impossible	Possible
Degraded V_{th} (V)	0.141530	0.080545	0.078727

PWCA, the proposed technique and greedy can select voltage levels using the real-time data. The difference between greedy and the proposed method is the voltage selection process. On the contrary to the greedy, the proposed method minimizes the following factor:

$$\alpha P \times \beta \Delta V_{th} \quad (11)$$

$$\alpha \propto \text{days have passed} \quad (12)$$

$$\beta \propto (10\text{yr} - \text{days have passed}) \quad (13)$$

where α and β are the weighted factors. P is power and ΔV_{th} is the predicted value of the increment in NBTI induced threshold voltage. During the early parts of lifetime, β is larger, NBTI stress is more weighted to the voltage selection. On the other hand, α becomes larger at the latter part of the target lifetime, thus power consumption is more weighted.

An example shown in Figure 1 presents characteristics of these methods. Greedy utilizes V_{dd} more aggressively during early parts, and V_{th} is increased faster. As a result, greedy should select more elevated voltages later. In contrast, both PWCA and the proposed method utilize V_{bs} more aggressively than greedy. At the latter parts, these two techniques utilize V_{dd} aggressively to take advantage of power consumption.

Table 1 shows the performance comparison of the proposed one with the previous works. The simulation scenario is equivalent to the example shown in Figure 1. The power values on Table 1 are average values of power consumption over the lifetime. Although PWCA has the lowest value, the proposed method has only 0.1% loss. The data reflection represents whether the algorithm recognizes the NBTI degradation and makes use. With PWCA method, data reflection is impossible, so the average power of PWCA should be held, even if NBTI stress intensity is reduced. Whereas the average power of greedy and the proposed method will be reduced if the reduced stress intensity is applied. The degraded V_{th} represents amount of V_{th} increase accumulated through the lifetime. As mentioned in section 2, the amount of V_{th} increase represents the NBTI progress state.

The simulation is organized with models described in section 2. The control policies are implemented in MATLAB. Power dissipation and hardware area are modeled with McPAT[10], power area timing modeling tool.

4. Conclusion

Voltage scaling technique is one of the most popular solutions for NBTI effect. In this paper, the voltage level selection process of voltage scaling technique is improved by considering stress intensity, and the average power consumption reduced roughly 12.9%. The pre-computed optimal voltage selection method (PWCA) induces excessive power consumption, because the stress intensity of real systems is not predictable. In addition, the average power loss is under 0.1% compared with PWCA. However, the proposed method does not incur the excessive power due to the fact that the monitored data reflection is possible.

Acknowledgement
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2012R1A2A1A03006255).

References

[1] A. Tiwari and J. Torrellas, "Facelift: Hiding and Slowing Down Aging in Multicores," *International Symposium on Microarchitecture*, IEEE, pp. 129-140, Nov. 2008.
[2] U. R. Karpuzcu, B. Greskamp, and J. Torrellas, "The BubbleWrap Many-Core: Popping Cores for Sequential Acceleration," *International Symposium on Microarchitecture*, IEEE, pp. 447-458, Dec. 2009.
[3] L. Zhang and R. P. Dick, "Scheduled Voltage Scaling for Increasing Lifetime in the Presence of NBTI," *ASP-DAC*, IEEE, pp. 492-497, Jan. 2009.
[4] Y. Lee, and T. Kim, "A Fine-Grained Technique of NBTI-Aware Voltage Scaling and Body Biasing for Standard Cell Based Designs," *ASP-DAC*, IEEE, pp. 603-608, Jan. 2011.
[5] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Adaptive Techniques for Overcoming Performance Degradation due to Aging in Digital Circuits," *ASP-DAC*, IEEE, pp. 284-289, Jan. 2009.
[6] E. Mintarno et al., "Self-Tuning for Maximized Lifetime Energy-Efficiency in the Presence of Circuit Aging," *TCAD*, IEEE, pp. 760-773 , May 2011.
[7] S. Bhardwaj et al., "Predictive Modeling of the NBTI Effect for Reliable Design," *CICC*, IEEE, pp. 189-192, Sept. 2006.
[8] W. Wang et al., "The Impact of NBTI Effect on Combinational Circuit: Modeling, Simulation, and Analysis," *TVLSI*, IEEE, pp. 173-183, Feb. 2010.
[9] R. Jejurikar et al, "Leakage Aware Dynamic Voltage Scaling for Real-Time Embedded Systems", *DAC*, IEEE, pp. 275-280, July, 2004.
[10] S. Li et al, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures", *International Symposium on Microarchitecture*, IEEE, pp. 469-480, Dec, 2009.