Scheduled Voltage Scaling for Increasing Lifetime in the Presence of NBTI

Lide Zhang
EECS Department
Northwestern University
Evanston, IL 60201, USA
lidezhang@u.northwestern.edu

Robert P. Dick
EECS Department
Northwestern University
Evanston, IL 60201, USA
dickrp@northwestern.edu

Li Shang from the University of Colorado, Boulder also participated in this project. He was omitted from the author list due to an error at the time of submission and it was against the policy of the conference to later add him to the author list.

Abstract—Negative Bias Temperature Instability (NBTI) is a leading reliability concern for integrated circuits (ICs). It gradually increases the threshold voltages of PMOS transistors, thereby increasing delay. We propose scheduled voltage scaling, a technique that gradually increases the operating voltage of the IC to compensate for NBTI-related performance degradation. Scheduled voltage scaling has the potential to increase IC lifetime by 46% relative to the conventional approach using guard banding for ICs fabricated using a 45 nm process.

I. INTRODUCTION

Integrated circuit (IC) process scaling is resulting in increasing vertical electric field. In addition, power density is increasing, which generally increases operating temperature. Together, these trends increase the rate of performance and reliability degradation due to negative bias temperature instability (NBTI). NBTI is a wear process that gradually increases the threshold voltage ($V_{th}$) of PMOS transistors and thus decreases circuit speed. An IC affected by NBTI will eventually fail if something is not done to compensate for this gradual performance degradation.

NBTI occurs due to the generation of interface traps at the Si–SiO$_2$ interface when a negative voltage is applied to the PMOS gate. When the traps accumulate hydrogen nuclei (H or H$_2$) diffusing towards the gate, the threshold voltage ($V_{th}$) of the transistor increases. When the stress is removed, no new interface traps are generated and hydrogen diffuses back. This leads to a decrease of the threshold voltage, $V_{th}$, which is called the recovery effect. For 70 nm technology, even when this recovery effect is considered, a 9.2% performance degradation after 10 years is predicted [1].

Several approaches have been proposed to alleviate NBTI-related degradation. The conventional solution to address decreasing performance is guard-banding in which the operating voltage or period is increased at design time in order to accommodate expected NBTI degradation. For instance, 10% $V_{DD}$ overhead may be required to tolerate 10% performance degradation due to an expected NBTI-related increase in threshold voltage ($V_{th}$) [2]. However, higher $V_{DD}$ directly increases the rate of NBTI degradation. It also results in higher power consumption and higher temperature, accelerating NBTI. High temperature and voltage may also accelerate other fault processes such as electromigration.

In this paper, we propose and evaluate scheduled voltage scaling for NBTI minimization and compensation. Instead of setting a fixed operating voltage guard band to compensate for potential degradation, the operating voltage is increased gradually during operation. This permits a $V_{DD}$ that is always lower than that required by guard-banding. Reduction in operating voltage or temperature each decrease wear rate due to NBTI [3]. Wear rate is directly decreased by the reduction in voltage, and indirectly decreased by the corresponding reduction in power consumption, and therefore temperature. Our method for determining an ideal voltage scaling schedule for NBTI minimization takes the accumulation of NBTI-related degradation into consideration. It also considers the impact of temperature on leakage power consumption, and therefore upon temperature. We demonstrate that scheduled continuous voltage scaling has the potential to increase the lifetime of a 45 nm process IC by 46% relative to that of guard-banding, with no increase in power consumption or temperature. In fact, power and temperature are decreased slightly. We also explain how to implement a discrete voltage scaling version of the technique.

The rest of the paper is organized as follows. Section II introduces related work on NBTI minimization. Section III presents the Reaction Diffusion based model of NBTI degradation and based upon the model, analyzes the overhead of guard banding. Section IV describes scheduled voltage scaling, assuming continuous voltages. Section V describes the impact of requiring discrete voltages on the technique and describe the implications for implementation. The rest of the paper is organized as follows. Section II introduces related work on NBTI minimization. Section III presents the Reaction Diffusion based model of NBTI degradation and based upon the model, analyzes the overhead of guard banding. Section IV
describes scheduled voltage scaling, assuming continuous voltages. Section V describes the impact of requiring discrete voltages on the technique and describe the implications for implementation.

II. RELATED WORK

People have proposed several techniques to tackle NBTI-related degradation while trading off different design metrics. In recent work, Srinivasan et al. proposed techniques for adding architecture-level redundancy for lifetime reliability that can also be applied to the NBTI problem [4]. However, the resulting area overhead and design complexity are undesirable. Frequency guard-banding is also widely used to permit continued operation in the presence of NBTI-related performance degradation. This technique consists of decreasing operating frequency to compensate for the potential speed degradation. Guard bands corresponding to 10–20% of the clock period are generally believed to be necessary to permit a 10 year operating lifespan. This unnecessarily sacrifices performance. Vattikonda, Wang, and Cao explored several techniques to minimize NBTI including $V_{DD}$ tuning, $V_{th}$ tuning and gate sizing [3]. However, setting a higher $V_{DD}$ yields more power and higher temperature, which may accelerate other fault processes, e.g., electromigration [4]. Likewise, decreasing $V_{th}$ increases leakage power consumption and therefore increases temperature. Moreover, tuning $V_{th}$ requires changes to oxide thickness or dopant density, complicating fabrication.

III. PREDICTIVE MODEL OF NBTI AND GUARD-BANDING

In this section, we first review the predictive NBTI-related $V_{th}$ degradation model based on the Reaction Diffusion mechanism [5]. The implications for lifetime, temperature, and power of guard-banding are then analyzed.

A. Overview of NBTI Degradation Model

A large number of previous NBTI models are developed based on the Reaction Diffusion mechanism [5]. The mechanism explains NBTI as a result of interaction of inversion layer holes with hydrogen-passivated Si atoms. The Reaction Diffusion based predictive model of NBTI considers the dependence of NBTI on the oxide thickness and the diffusing species ($H$ or $H_2$). It also provides a closed-form expression for the threshold voltage change ($\Delta V_{th}$). The static NBTI at time $t$ can be expressed as

$$\Delta V_{th}(t) = A[(1 + \delta)t_{ox} + \sqrt{Ct}]^{2n}$$  \hspace{1cm} (1)

where $n = 1/6$ for a $H_2$ diffusion based model and $n = 1/4$ for a $H$ based model. In this work, the focus is $H_2$ based diffusion. $\delta$ is a curve fitting constant and can be ignored for long-term effects. $A$ and $C$ are two time-independent coefficients. $A$ is linearly proportional to the hole density and has an exponential dependence on the electric field ($E_{ox} = (V_{GS} - V_{th})$). $C$ is exponentially dependent on the temperature $T$.

$$A = \left(\frac{q t_{ox}}{e}\right)^{\frac{1}{2}} K^2 C_{ox} (V_{gs} - V_{th}) \left(e^{\frac{E_{ox}}{kT}}\right)^{\frac{1}{2}}$$  \hspace{1cm} (2)

$$C = T_0^{-1} \cdot \exp\left(-E_a/kT\right)$$  \hspace{1cm} (3)

$q$ is the electron charge, $k$ is Boltzmann’s constant, $C_{ox}$ is the oxide capacitance per unit area, and $E_a \approx 0.49 \text{eV}$ when the diffusing species is $H_2$. Equation 1 can be simplified by eliminating $(1 + \delta)t_{ox}$ because $(1 + \delta)t_{ox} \ll \sqrt{C(t - t_0)}$ for long-term effects. Thus $\Delta V_{th}$ can be expressed as a function of $t$ and $A$, i.e., $\Delta V_{th} = A(C \cdot t)^n$. This yields [5]

$$\Delta V_{th} = \left(\sqrt{K^2_{2} \cdot T_{clk} \cdot \alpha/ \left(1 - \beta_1^{1/2n}\right)}\right)^{2n}$$  \hspace{1cm} (4)

where

$$\beta_1 = 1 - \frac{2\xi_1 \cdot t_{ox} + \sqrt{\xi_2 \cdot C \cdot (1 - \alpha) \cdot T_{clk}}}{2t_{ox} + \sqrt{C \cdot t}}$$  \hspace{1cm} (5)

$\alpha$ is the signal’s duty cycle; $\xi_1$ and $\xi_2$ are two constants. Note that the degradation rate is a weak function of duty cycle. Hence, although degradation depends on duty cycle, when considering various topologies of circuit, different PMOS transistors might be stressed with different duty cycles. Moreover, decreasing the stress time of one transistor may increase the stress time of following transistors and thus makes the whole circuit less sensitive to the duty cycle. Therefore, analysis based on worst-case stress will generally not result in a significant over-estimation of degradation.

B. Guard-Banding

Guard-banding is a technique to tolerate the degradation by degrading circuit frequency, threshold voltage, gate sizes, and increasing operating voltage. Among them $V_{DD}$ tuning is currently dominant [3]. $V_{DD}$ tuning fixes the operating voltage at a higher value than necessary to guarantee correct operation even after NBTI-related performance degradation. However, the timing slack preserved for compensation for lifetime NBTI degradation not only slows down the circuit performance in early life, but also consumes more energy and generates more heat during operation. Analysis based on the predictive NBTI model shows that a voltage overhead of approximately 14.5% is required to compensate for the degradation in 10 years. As a result, the average temperature is increased by 5.63% and the power consumption is increased by 30%. Both higher temperature and higher $V_{DD}$ result in faster NBTI degradation.

Instead of using a higher static voltage, scheduled voltage scaling always uses the lowest voltage required to maintain adequate performance. The reduced power and temperature resulting from using a lower voltage has the added benefit of extending lifetime. One question remains: how should scheduled voltage scaling be modeled and analyzed in order to determine an optimal voltage change schedule?
IV. SCHEDULED VOLTAGE SCALING

In this section, we describe modeling and analysis of scheduled voltage scaling, assuming the availability of continuous voltage levels. We will consider discrete voltage levels in the next section. In the latter part of this section, simulation results showing the impact of the proposed approach on lifetime and power consumption.

A. Problem Definition and Analysis

To maintain the performance and therefore reliability of the circuit during its entire operating lifetime, $V_{DD}$ must be gradually increased to compensate for NBTI-related degradation. Therefore, it is necessary to decide the optimal voltage in each time interval. Increasing $V_{DD}$ increases temperature, which in turn increases degradation rate. It is also necessary to model NBTI degradation under changing $V_{DD}$ and take temperature into consideration.

Increased $V_{DD}$ increases NBTI degradation rate, as shown in Figure 1. In this example, $V_{DD}$ is initially low (1 V) and is set high (1.2 V) at time $t_1$. Functions $\Delta V_{th1}(V_{DD})$ and $\Delta V_{th2}(V_{DD})$ model the different degradation rates of $\Delta V_{th}$ corresponding to voltages 1 V and 1.2 V. Due to the accumulation of wear, the degradation in threshold voltage before and after the voltage transition should be the same, i.e.,

$$\Delta V_{th1}(t_1) = \Delta V_{th2}(t_2)$$

where $t_2$ is the equivalent age under $V_{DD2}$. Therefore, whenever $V_{DD}$ is adjusted, it is necessary to switch from one wear-out function to another. The current time also needs to be adjusted to a value in the new time scale that corresponds to the same amount of $V_{th}$ degradation. This new value of $t$ is defined as equivalent age. Consider the $(i+1)$th change in voltage, the degradation after $i+1$ periods can be expressed as

$$\Delta V_{th}(i+1) = A(i+1) \cdot [C(i+1) \cdot (t'+t)]^n$$

where $t'$ is the equivalent age of $\Delta V_{th}(i)$ at $V_{DD(i+1)}$ and $t$ is the operating time at voltage $V_{DD(i+1)}$. Both $A$ and $C$

can be expressed as functions of $V_{DD(i+1)}$. After getting the equivalent age $t'$, the equation can be expressed as

$$\Delta V_{th}(i+1) = [\Delta V_{th}(i) \cdot C(i+1) \cdot A(i+1) \cdot t']^n$$

Based on the above equation, $\Delta V_{th}(i+1)$ is expressed as a function of the previous time period $\Delta V_{th}(i)$ and the operating voltage during the current time period.

As indicated in Equations 1 and 2, NBTI degradation depends exponentially on temperature. Moreover, due to the super-linear relationship between temperature and $V_{DD}$, when voltage increases, temperature increases, accelerating degradation and thus the requirement to again increase voltage. Hence, the impact of temperature cannot be neglected. The relationship between temperature and power consumption can be expressed as follows [7]:

$$T_{processor} = T_A + P \cdot R_{thermal}$$

where $T_A$ is the ambient temperature, $P$ is the total power including both dynamic power and leakage power, and $R_{thermal}$ is the thermal resistance from the processor to the ambient.

Leakage power accounts for 40% of the power consumption in modern high-performance microprocessors [6], and most of this is due to subthreshold leakage. Leakage power can be approximated as follows [7]:

$$P_{leakage} = V_{DD} \cdot I_{leakage}$$

$$= V_{DD} \cdot A_s \cdot \frac{W}{L} \cdot \left(\frac{kT}{q}\right)^2 \cdot e^{\frac{Q(V_{DD} - V_{th})}{nkT}}$$

$A_s$ is a technology-dependent constant, $n (=1.5)$ is the subthreshold swing coefficient for the transistor [8], and $T$ is the temperature. Equations 9 and 10 indicate that leakage power and temperature are interdependent. We account for this by iteratively computing leakage power and temperature until convergence.

At this point, we have all the necessary models to determine an optimal voltage schedule. According to the alpha power law [9], the delay given $V_{DDi}$ and $V_{thi}$ is

$$d_i = \frac{K \cdot V_{DDi}}{(V_{DDi} - V_{thi})^n}$$
where $K$ is a delay constant and $\alpha = 2$. We then define the following variable and constant. $d_{\text{initial}}$ is the constant performance constraint. $V_{DD}(t)$ is the optimal voltage at time $t$, and $V_{\text{th}}(t, V_{DD}(t))$ is the threshold voltage given $V_{DD}(t)$ at time $t$. To maintain the same performance over years, solving the following equation

$$d_{\text{initial}} = \frac{K \cdot V_{DD}(t)}{(V_{DD}(t) - V_{\text{th}}(t, V_{DD}(t)))^\alpha} \quad (13)$$

yields the optimal voltage schedule. Figure 2 shows the continuous voltage scaling scheme for a 45 nm process technology integrated circuit. Using guard-banding, the necessary voltage is 1.143 V and the delay increases by 16% relative to its initial value. Scheduled voltage scaling gradually changes the voltage from 1 V to 1.135 V in 10 years. The voltage necessary for scheduled voltage scaling is always smaller than the voltage for guard-banding.

**B. Simulation Results**

Before applying the proposed technique, we first implement the simpler model without considering the influence of varying temperature. Assuming a fixed temperature when modeling voltage scaling simplifies the model to a great extent. However, analysis shows that an NBTI model that does not consider temperature change suggests using a guard band of 0.1223 V while an NBTI model that accounts for time-varying temperature suggests a guard band of 0.147 V, each for a ten-year operating lifetime. Neglecting temperature results in 16.83% error in the approximation of guard band voltage, which would in turn introduce 52% error when estimating lifetime. Therefore, considering the influence of temperature is necessary.

In this section, we compare the lifetime permitted by scheduled voltage scaling with that permitted by voltage guard banding. Let us first compare the lifetimes of both techniques given the same voltage constraints. We assume the $V_{DD}$ eventually reached when using guard-banding is the maximum voltage tolerated by the IC and both cases have the same performance (frequency) requirement. To meet the performance constraint, guard-banding works at the maximum voltage and fails when the delay reaches its constraint, while voltage scaling maintains constant delay and operates until $V_{DD}$ reaches the voltage constraint. $V_{DD}$ in the proposed technique never exceeds $V_{DD}$ in guard-banding during the circuit’s lifetime. Table I demonstrates the lifetime improvements of scheduled voltage scaling relative to guard banding for a number of voltage constraints. For example, given a maximum voltage of 1.147 V, the lifetime of the proposed technique is 14.6 years, i.e., an increase of 46% comparing to guard-banding.

The necessary maximum voltage for scheduled voltage scaling is smaller than that using guard-band and thus saves power. Table II summarizes the improvement of power and maximum voltage compared to guard-banding within the same operating period, when both have the same lifetime constraint. We consider 45 nm process technology. For instance, in a 10 year period, the maximum voltage is reduced by 7.62% and thus the maximum power consumption is 4.16% lower.

Figure 3 shows the influence of process technologies on the benefit of scheduled voltage scaling relative to guard banding. The benefit increases with the original lifetime using guard-banding. This is because the higher the temperature produced by guard-bANDING, the more rapidly PMOS transistor performance degrades. Note that the simulation results are based on a signal with 100% duty cycle, which is the worst case for NBTI degradation.

![Fig. 3. Increase in lifetime due to scheduled voltage scaling](image)
V. DISCRETE LEVELS VOLTAGE SCALING

The previous section introduces the theoretical continuous voltage scaling model. To implement the technique in practice, discretization is necessary. This section describes a method of implementing this technique using a voltage regulator that supports the sets of discrete voltages. In addition, we describe the influence of discretization on integrated circuit lifetime, power consumption, and voltage regulator design. Finally, the control circuit of the voltage regulator is discussed.

A. Circuit-Level Implementation Issue

The voltage range used in scheduled voltage scaling is much smaller than that normally used in voltage regulators designed for dynamic voltage scaling. Therefore, modification of existing regulators is necessary to support the small voltage steps indicated in Table I. LTC3447 is a high-efficiency monolithic synchronous current mode buck regulator [10]. Using an I²C interface, the output voltage can be set between 0.69 V and 2.05 V with 21.6 mV resolution using an internal 6-bit Digital to Analog Converter (DAC). By replacing the DAC with one having a smaller voltage step, i.e., a calibration DAC, the voltage step of the regulator can be decreased to 1.37 mV [11].

Supporting small voltage steps is necessary, but not sufficient, for scheduled voltage scaling. As many ICs have their own voltage regulators for use in dynamic voltage and frequency scaling (DVFS), one challenge is to integrate the NBTI-aware voltage levels into the existing set of voltages supported by the regulator. As shown in Figure 4, the newly added voltage levels can be seen as discrete voltage levels clustering below the original voltages. $V_0$, $V_1$, and $V_2$ are the original voltages. The other lines indicate scheduled voltage scaling levels. As a result, a voltage regulator providing non-uniform distributed voltage levels is required. To solve this, a voltage regulator with higher resolution DAC is used. More voltage levels can be provided, but the resulting set of voltages is unevenly distributed. This can be used to improve the design of the DAC. Figure 5 shows the schematic for a $n$-bit DAC. To simplify the design, we remove all the structure attached to the bit lines that need not to change. For instance, the MOSFET, the corresponding resistor, and the inverters are all removed from bit lines associated with static values. This simplifies the design to some extent.

B. Impact of Discretization

Given the trade off between sub-optimality and complexity that depends on the number of voltage levels, it is necessary to choose the number based on the marginal improvement in lifetime. Table III shows the sensitivity of lifetime improvement to number of voltages. Increasing from 3 levels to 4 levels increases lifetime by 139%, while increasing from 20 levels to 21 levels only yields a 0.97% improvement. If increases to the number of voltage levels yielding less than 5% improvement in lifetime are prevented, 10 is voltages levels are implied. Figure 6 describes the 6 levels discretized voltage scheme corresponding with its scheduled switching time. During normal operation, voltage switching only occurs when the performance constraint cannot be satisfied due to NBTI.

C. Degradation Detection

At this point, a fine-grained voltage regulator and a discrete voltage scheme have been presented. The last question is when and how we should switch from a lower voltage to a higher one. To tackle this problem, multiple approaches are proposed here for different applications and they can be categorized into real-time technique and static technique.

For ICs that do not use DVS, we can use static technique to predetermine the voltage scheme and the switching points can be pre-calculated according to the estimation of the degradation. Timing information can be stored in non-volatile memory. To make the circuit to be aware of the current degradation level, a counter may be added to record the age

<table>
<thead>
<tr>
<th>Number of voltage levels</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit ($\times$)</td>
<td>1.046</td>
<td>1.11</td>
<td>1.162</td>
<td>1.203</td>
<td>1.29</td>
<td>1.325</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>139.1</td>
<td>47.3</td>
<td>25.3</td>
<td>14.8</td>
<td>4.5</td>
<td>0.97</td>
</tr>
</tbody>
</table>
of the circuit. By comparing the timing table with the age of the circuit, the control circuit can determine whether to switch to a higher level. The static technique does not require any significant changes to the hardware.

For chips with existing DVFS scheme, the unpredictable changes in voltage levels make the estimation of degradation due to NBTI at design time impossible. Hence, two real-time techniques may be used to either implicitly or explicitly track degradation. The software-based technique is a program recording the voltage changing scheme due to DVFS. Both the current degradation and delay can be estimated with the accumulated wear rate model based on the dynamically-changing voltage. When the estimated delay reaches the maximum delay allowed by the current voltage, we switch to a higher voltage level. The accumulated current degradation is stored in non-volatile memory like EEPROM off chip, which is updated periodically. As the non-volatile memory cannot be accessed immediately after booting. Therefore, to avoid timing violations, the highest voltage level should always be chosen when booting. After the current degradation is correctly estimated, a lower voltage can be set. As is the case for the static technique, this approach will not induce any significant change in hardware.

Another approach is hardware-based technique that tracks the changing delay directly. A delay sensor is placed in the critical path of the circuit [12]. To determine the switching time, several comparators, each connected to a counter, are attached to the output of the delay sensor. One comparator corresponds to one voltage level and the reference voltage of it is determined by the maximum delay under the corresponding voltage level. By checking the value of the counters, we can determine the age of the circuit and therefore decide the time to switch to a higher voltage level. The overhead of the delay sensor is less than 2% in delay and less than 4% in area.

VI. CONCLUSION

We have proposed a technique called scheduled voltage scaling to optimize IC lifetime in the presence of NBTI. This technique has the potential to increase IC lifetime by 46% relative to guard-banding for 45 nm process technology. We have also evaluated the effect of using a discrete voltage scheme permitting implementation with only a minor change to current voltage regulator design styles. We evaluated the marginal improvement of adding additional voltage levels. The discrete voltage scheme using 10 voltage levels increases 45 nm process technology IC lifetime by 32.5% relative to guard-banding.

REFERENCES

[10] Linear Technology, +FC Controllable Buck Regulator in 3mm × 3mm DFN.