Built-In Proactive Tuning System for Circuit Aging Resilience

Nimay Shah\textsuperscript{1}, Rupak Samanta\textsuperscript{1}, Ming Zhang\textsuperscript{2}, Jiang Hu\textsuperscript{1}, Duncan Walker\textsuperscript{3}

\textsuperscript{1}Dept. of ECE, Texas A&M University, College Station
\textsuperscript{2}SoC Enabling Group, Intel
\textsuperscript{3}Dept. of Computer Science, Texas A&M University, College Station

E-mail: nimay\_shah@tamu.edu, rupak9@tamu.edu, ming.y.zhang@intel.com, jianghu@ece.tamu.edu, walker@cs.tamu.edu

Abstract

VLSI circuits in nanometer VLSI technology experience significant aging effects, which are embodied by performance degradation over operation time. Although this degradation can be compensated by over-design, it induces remarkable power overhead which is undesirable in tightly power-constrained designs. Dynamic voltage scaling (DVS) is a more power-efficient approach. However, its coarse granularity implies difficulty in handling fine-grained variations in the aging effects. We propose a Built-In Proactive Tuning (BIPT) system that allows each circuit block to autonomously tune its performance according to its own degree of aging. The BIPT system is validated through SPICE simulations on benchmark circuits with consideration of NBTI effect. The experimental results indicate that the proposed BIPT system leads to about 45\% less power than the approach of over-design while maintaining the same performance. Compared to DVS, BIPT can achieve the same aging resilience with about 30\% less power dissipation.

1. Introduction

As VLSI technology scales to nanometer regime, circuit aging effects, such as NBTI (Negative Bias Temperature Instability) and HCI (Hot Carrier Injection) \cite{10} become prominent. NBTI manifests itself by degradation of PMOS threshold voltage \cite{6, 10} whereas HCI results in threshold voltage increase in mostly NMOS transistors. When technology scales from 180nm to 65nm, the MTTF (Mean Time To Failure) of processors due to aging effects is reduced by about 76\% \cite{10}. That is, if a chip would have previously lasted for 10 years, now it can perform well for about 2 years. Therefore, it becomes increasingly imperative to address the aging effect in chip design.

To handle the circuit aging problem, a common approach is to over-size transistors such that the aged performance can still meet specifications \cite{6}. This approach is able to extend chip lifetime under the aging effect. However, it inevitably increases circuit power dissipation and therefore hits another wall of nanometer integrated circuit design – the increasingly tight power constraint. Over-sized transistors usually imply unnecessarily large timing slack and therefore wasteful power dissipation during the initial lifetime of a circuit. Alternatively, architectural approaches \cite{9, 10} are suggested for mitigating the aging problem. One technique is architectural-level adaptation \cite{9, 10} such as DVS (Dynamic Voltage Scaling). For instance, a chip can operate at relatively low supply voltage level when new and switch to higher supply voltage level when it gets aged. Such adaptation can avoid the wasteful power as compared to the oversized transistor approach. However, this is a coarse-grained technique – the supply voltage level is usually fixed for major partitions of the chip, if not across the entire chip. In general, the aging effects vary among different components of a circuit. In order to ensure the performance of an entire chip, the DVS must be performed according to the worst transistor aging. That is, strong aging of only 1\% percent transistors may require the
chip-level supply voltage increase although the other 99% transistors have very minor aging induced degradation.

In this paper, we propose a Built-In Proactive Tuning (BIPT) system to mitigate the aging problem in a power-efficient manner. This system includes a canary circuit which can generate predictive warning signals for performance degradation. According to the warning signal, circuit speed is tuned through body bias such that the performance degradation is compensated. The proactive tuning is performed offline, at power-on of the chip or periodically. Since aging is a slow change with time constant of weeks/months, periodic tuning of once in a few days is sufficient to capture the change. The offline tuning has the advantage of allowing relatively easy control on input vectors. When detecting performance degradation or circuit delay variation, one must consider the delay uncertainty due to different input vectors. Even if there is no warning signal for certain input vectors, there is still risk of delay errors under other input vectors. Therefore, we include a Test Pattern Generator (TPG) in the system in order to have large input vector coverage. TPG is usually a part of Built-in Self Test (BIST) hardware; thus, if a chip already has BIST circuit, TPG does not cause extra overhead. The proposed BIPT system has the following advantages:

- It can be applied at circuit block level instead of the chip level architectural approach [9, 10]. In other words, each block can be tuned according to its own degree of aging. Evidently, the finer granularity control allows improved power efficiency.
- Its performance degradation detection is obtained from the actual operating circuit as opposed to replica circuit in other adaptive design methods [12]. Since the detection is more direct, it is more reliable. Using TPG further improves the reliability of the detection.
- Its proactive nature can avoid the complex error correction schemes in retroactive systems [3, 4]. The retroactive systems rely on pipeline flush [4] or instruction replay [3] and therefore are restricted to processor designs. In contrast, our system can be applied to both processors and general sequential circuits.

Existing approaches have one or two of the above advantages, but none of them have all to the best of our knowledge. The work of [12] is a block level adaptive body bias technique. However, its delay variation detection is obtained from replica circuits which often have discrepancy from the actual operating circuits. The Razor based techniques [3, 4] use direct variation detection, but they rely on complex error correction method and are restricted to processor designs. Another retroactive method [7] is mainly targeted for fast variations such as voltage variations and hence complements our work. The canary circuit based predictive detection is proposed in [8]. However, it is applied with online tuning which suffers from delay uncertainty due to different input vectors. The recent work of [2] focuses on only the aging detection instead of an overall tuning system. Actually, the detection method in [2] can be easily adopted in our tuning system.

The BIPT system is validated through SPICE simulations on benchmark circuits with consideration of NBTI effect. Even with consideration of overhead due to TPG, canary and control circuit, the proposed BIPT system can lead to about 45% less power than the over-design approach while maintaining the same performance. Compared to DVS, BIPT can achieve the same aging resilience with about 30% less power dissipation.
2. Built-In Proactive Tuning System

2.1. Overview

The Built-In Proactive Tuning (BIPT) System consists of the existing main circuit augmented with a Test Pattern Generator (TPG), Body Bias Circuitry, Canary Circuit and Control circuit. Figure 1 shows these blocks and the corresponding interface signals.

![Figure 1. Overview of the proposed built-in proactive tuning system](image)

At power-on or periodically, the BIPT system can launch test vectors from TPG and then tune circuit body voltage according to the observations from the canary circuit. Canary circuit plays the role of predicting aging-induced performance degradations (more details in Section 2.2). A Warning signal is generated by the canary flip-flops when the timing constraint is tight on one or more of the few critical paths where these are inserted. The top-level warning signal is the \textit{OR} of all the individual canary flip-flop warning signals. The Linear Feedback Shift Register (LFSR) \cite{1} is implemented as a pseudo-random test pattern generator which applies these random patterns when offline test is in progress. It is triggered by the \textit{preset} signal from the control block. The control block monitors the status of all the blocks and issues control signals. \textit{PON} is the power-on-reset signal which is an active high reset signal issued on start-up and basically triggers the offline test. \textit{Offline test} is an active high signal indicating that offline test is in progress. The most critical activity performed by the control block is to monitor the warning signal from the canary circuit. Based on this signal, it appropriately sets the body bias to selective gates on the critical paths of the main circuit via the bias level signal passed to the body bias block. This interface and the body bias block are modeled as in \cite{12}. The body bias is adaptive to the circuit state: the circuit automatically selects from 4 available options of forward body bias using a counter – decoder based scheme. This has been described in detail in section 2.3.
2.2 Canary Circuit

The canary circuit [8] is for detecting aging-induced performance degradation in a predictive manner. As shown in figure 2, a canary circuit consists of two flip-flops; a main FF and a canary FF. The main FF gets the direct input and the canary FF which serves as the checker part gets the input through a delay buffer. This delay in the input reaching the two flops serves as the guard band for error detection. The outputs from these flops are fed to an xor gate which functions as a comparator, outputting 1 when these are different and thereby predicting the occurrence of an error. Some advanced designs of canary circuits are proposed in [2, 13].

![Figure 2. Canary circuit](image)

Canary circuit is a typical case design alternative of Razor [4]. However, in contrast to Razor, which delivers a delayed system clock to the checker part (shadow FF), canary circuit delivers a delayed input signal to the checker part (canary FF). This simplifies the clock tree synthesis and routing as there is just one system clock now. Also, the delay buffer placed before the canary flip-flop always has a positive delay, even if affected by aging, which makes the canary flip-flop recover from variation induced effects by itself. Canary circuit also predicts timing errors rather than detecting them afterwards. The predictive warning allows the user to take preventive measures before the timing violation actually occurs and thus the system does not run into any corrupt data states, except for errors that cannot be predicted such as single event upset (SEU) errors. However, aging induced timing violations can be predicted effectively by architectures such as canary circuit [2].

2.3. Test Pattern Generator and Control Circuit

Figure 3 shows the gate level implementation of the control circuit. *Finish* signal going high indicates the completion of offline testing, *PON* is the power-on-reset signal, *Warning* is the timing error prediction signal from the canary circuit and *Preset* is the active low signal to set the flip-flops in the LFSR to high state on power-on-reset. The preset generation circuit is shown in the dotted box in figure 3. The initial states of all the flip-flops in the LFSR on power-on-reset is ‘1’, thus the starting seed for the LFSR is all 1’s. The LFSR shown in figure 3 is a 12-bit LFSR; it implements a primitive polynomial to generate 4095 patterns ($2^n-1$; $n=12$) before returning back to the initial state of all 1’s. The outputs of the flip-flops in the LFSR are fed to a scan chain through a mux-d connection. These connections are omitted in figure 3 for clarity.
The control circuit is triggered by the power-on-reset signal (PON), which remains high for one cycle on each power-on-reset of the chip. On each power-on-reset, the offline test signal triggers the offline testing. Generation of the offline test signal is shown in the box in Figure 3. The offline test signal is the input for preset generation circuit that presets the flip-flops in the LFSR to high, the initial seed for the test patterns in the LFSR. The Finish signal is generated by the circuit shown in Figure 4(a). Its first stage consists of a 12-input AND gate and the second stage consists of a 2-input Muller-C gate. Muller-C gate is an AND gate for events i.e., it produces a high output when all the inputs are high and goes low only when all the inputs transition to low state. The description about Muller-C gates can be found in [16]. As shown in Figure 4(a), the outputs of the flip-flops in the LFSR are connected to a 12-input AND gate. The output of this AND gate and PON feed to a 2-input Muller-C gate to produce the finish signal. On every power-on-reset, the flip-flops of the LFSR are preset to 1, thus the output of the AND gate rises high. Since PON is active high, PON is low at startup and thus finish stays at 0 initially. PON stays high for one clock cycle and then goes to low. When all the 4095 test patterns have been generated, the output of the AND gate goes high again and since PON is also high; finish goes high indicating the completion of offline testing. After finish goes to high, at the next clock edge, the output of the AND gate goes low due to a pattern other than all ones. However, the finish signal still stays high because of the property of the Muller-C gate to hold the previous value until both the inputs transition to the same value. In this case, although the output of the AND gate goes low, since PON is still high, finish stays high.

The possible timing violations in the critical paths i.e., the paths that are affected due to aging are predicted by the warning signal from the canary flip-flops. The critical paths that
are affected by aging need to be corrected by application of suitable forward body bias. Since, the body bias generation circuit takes some time to apply correct bias to the devices on these critical paths, the LFSR needs to be stalled. In our approach, we stall the clock to the LFSR by using gated clock circuitry shown in figure 4(b). In figure 4(b), the circuit can stall the clock for one clock cycle, which is sufficient for us to change the body bias of the devices on the critical paths. However, if we need more time then the clock can be stalled for a longer period of time using cascaded Muller-C gates in figure 4(b).

**Figure 4(a). Finish signal generator**  
**Figure 4(b). Gated clock circuit**

**Figure 5. Generation of control signal to body bias block**

Figure 5 shows the body-bias generation logic. As shown, it consists of a 2-bit up-counter, made up of two flip-flops. The four possible states of the counter translate into fours possible body bias levels to choose from. Level 0 is the no-bias condition; levels 1-3 are in the increasing order of the forward body biases. As stated earlier, we deal with aging degradation which monotonically degrades the circuit performance with time; thus, forward body bias or reduced reverse body bias is necessary to restore the circuit performance. The up-counter counts upward (increases forward bias / reduces reverse bias) when a warning signal is generated by the canary circuit. It counts upward till it reaches the highest forward body bias state / least reverse body bias (binary 11 in our case) and freezes in that state. We implement a four state counter as few forward body bias levels are sufficient for the circuits under consideration. However, larger number of forward bias levels can be generated by adding extra flip-flops in the body bias generation circuit. The outputs Q1 and Q2 of the counter go to a 2-to-4 decoder. The decoder outputs are inputs to the body bias circuitry which is implemented as in [12] and enable the appropriate body bias option.
3. Experiment Setup and Results

The experiments for offline testing are performed on ISCAS89 sequential benchmarks: s526 and s832. First, we augment these circuits with BIPT hardware. To do this, we determine the critical paths in these circuits by using a static timing analyzer written in C. The gate libraries needed for the static timer are characterized in HSPICE for 90nm model card from BPTM [http://www-device.eecs.Berkeley.edu/~ptm/]. The flip-flops at the output of the critical paths are replaced by canary circuits whose structure and operation is described in section 2.2 and in detail in [8]. Once the canary circuits are determined, we traverse the path from input of the canary FF in a breadth first manner till we reach either a flip-flop or a primary input. We replace the flip-flops by mux-d scan flops and add extra scan flip-flops for the primary inputs. The scan flip-flop has two inputs, one input is connected to the input of the original flip-flop and the other input is connected to the output of the LFSR. A scan-enable signal is used to select between the two inputs. Finish signal serves as the scan-enable for scan flip-flops in our design. It can as well be a user-defined input. The characteristics of the benchmarks pre- and post-BIPT processing are shown in Table 1. Column 3 shows the number of flip-flops originally in the design and column 7 shows the number of these flip-flops replaced by canary flops respectively. Column 6 shows the number of mux-d scan flops inserted in the design. To validate the BIPT system on these benchmarks, we consider the effect of NBTI induced PMOS threshold voltage ($V_t$) degradation in these circuits. Our simulations take into account the effect of both nominal $V_t$ degradation and temporal variations in $V_t$ degradation; using models as described in [5]. The other important task is to set the clock period for simulations and thus set the target performance for both the benchmarks. The clock period for the simulations is determined by applying a pre-defined $V_{dd}$ to just the main circuit (without BIPT hardware). For this $V_{dd}$, we run simulations to find out the nominal clock period such that no error occurs during offline testing. We add a safety margin of 15% to this period and the resulting clock period becomes clock period for our simulations. For a $V_{dd}$ of 1.15V, this final value is found to be 480ps (2.08 GHz) and 600ps (1.67 GHz) for s526 and s832 respectively. Since a circuit with BIPT hardware, doesn’t need to operate with high safety margins, we set the clock period to be 480ps (600ps) and for this clock period we determine minimum $V_{dd}$ such that no error occurs during offline testing. For both the benchmarks, $V_{dd}$ is set to 0.925v for BIPT.

Table 1. Characteristics of ISCAS ’89 benchmarks under consideration: pre-BIPT processing and post-BIPT processing

<table>
<thead>
<tr>
<th>ISCAS’89 Benchmark</th>
<th>No of Gates</th>
<th>No of flip-flops (FF) (pre-BIPT processing)</th>
<th>No of Primary Inputs</th>
<th>No of Primary Outputs</th>
<th>No of mux-d scan-flops (post-BIPT processing)</th>
<th>No of FF replaced by canary FF (post-BIPT processing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s526</td>
<td>193</td>
<td>21</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>s832</td>
<td>262</td>
<td>5</td>
<td>18</td>
<td>19</td>
<td>12</td>
<td>2</td>
</tr>
</tbody>
</table>

To evaluate the effectiveness of BIPT approach, we carry out two sets of simulations in HSPICE at 100° C: (a) Deterministic Simulations and (b) Statistical Simulations. For deterministic simulations, simulations are carried out for 0%, 5% and 10% of NBTI induced deterministic $V_t$ degradation. We compare the total operating power consumed by BIPT scheme with the over-designed case as the baseline case. The power estimation of BIPT system includes power dissipation due to the TPG, canary circuit and control circuit. The over-design implemented here is a conservative scaling of $V_{dd}$ level. In particular, the $V_{dd}$ for the over-designed case is set such that it does not cause timing violations and meets the...
performance targets at 10% $V_t$ degradation as well. This value is found to be 1.2V for both s526 and s832 for 2.08 GHz and 1.67 GHz respectively. On the other hand, BIPT scheme allows for typical case circuit design, and adapts to the degradation of the circuit during its lifetime. Thus, the operating voltage is kept at 0.925V for BIPT simulations. Figure 6 plots the power consumed for deterministic simulations for s526 and s832. From the simulation results, we can observe that, on an average, BIPT scheme leads to power savings of 45% compared to the over-designed case.

![Power consumption for Over-Designed vs BIPT schemes, Deterministic Simulations](image1)

**Figure 6. Power consumption for deterministic simulations**

![Power consumption for DVS vs BIPT schemes, Statistical Simulations](image2)

**Figure 7. Power consumption for statistical simulations considering the temporal variations of NBTI effect**

For statistical simulations, we take into account the statistical component of NBTI degradation over and above the nominal $V_t$ degradation in lifetime $V_t$ degradation, which takes into account the statistical variation in the underlying process causing $V_t$ degradation [5]. We model the statistical $V_t$ degradation as a Poisson random variable. We compare the power consumed by BIPT scheme with Dynamic Voltage Scaling (DVS) scheme. Thus, the dynamic voltage scheme serves as the baseline for statistical simulations. The simulations are carried out for statistical $V_t$ variation over 2%, 5% and 10% of nominal value. The $V_{dd}$ values for DVS are selected such that in each nominal case, the circuit is ensured to work for the
worst statistical variation. Thus, for a transistor whose $V_t$ is degraded by 5% (statistical variation component) over and above the 2% nominal degradation, $V_{dd}$ is selected such that the circuit would still work without any timing violations if all transistors in the circuit were similarly affected. The operating voltages for the DVS schemes are found to be 1.15V, 1.2V and 1.25V for 2%, 5% and 10% degradations respectively. The operating voltage for BIPT case still remains at 0.925V. Figure 7 plots the power consumed for statistical simulations for s526 and s832. From the experimental results, we can observe that, on an average, BIPT scheme leads to power savings of 30% compared to the dynamic voltage scaling approach. Power saving here is less than the over-design case because dynamic voltage scaling scheme is an improvement over the over-designed approach. The power for DVS methodology increases as $V_t$ degradation increases because of the fact that the voltage supply is varied keeping in mind the most degraded transistor.

4. Conclusions

In this paper, we propose a Built-In Proactive Tuning system that allows VLSI circuits to autonomously compensate aging-induced performance degradations. Due to its adaptive nature, BIPT is power-efficient and uses about 45% less power than over-design based aging compensation. Since it is a middle-grained approach, it can achieve 30% power reduction compared to the coarse-grained DVS method.

References