Adaptive Error-Prediction Flip-flop for Performance Failure Prediction with Aging Sensors

C. V. Martins, J. Semião
Univ. of Algarve / INESC-ID Lisbon
Faro, Portugal
{cvmartins, jsemiao}@ualg.pt

J. C. Vazquez, V. Champac
INAOE
Puebla, Mexico
{jcvazquez, champac}@inaoep.mx

M. Santos, I. C. Teixeira, J. P. Teixeira
INESC-ID Lisbon / IST-UTL
Lisbon, Portugal
{marcelino.santos, isabel.teixeira, paulo.teixeira}@ist.utl.pt

Abstract—This paper presents a new approach on aging sensors for synchronous digital circuits. An adaptive error-prediction flip-flop architecture with built-in aging sensor is proposed, performing on-line monitoring of long-term performance degradation of CMOS digital systems. The main advantage is that the sensor’s performance degradation works in favor of the predictive error detection. The sensor is out of the signal path. Performance error prediction is implemented by the detection of late transitions at flip-flop data input, caused by aging (namely, due to NBTI), or to physical defects activated by long lifetime operation. Such errors must not occur in safety-critical systems (automotive, health, space). A sensor insertion algorithm is also proposed, to selectively insert them in key locations in the design. Sensors can be always active or at pre-defined states. Simulation results are presented for a balanced pipeline multiplier in 65 nm CMOS technology, using Berkeley Predictive Technology Models (PTM). It is shown that the impact of aging degradation and/or PVT (Process, power supply Voltage and Temperature) variations on the sensor enhance error prediction.

Keywords: aging sensor; performance failure prediction; NBTI; delay insertion.

I. INTRODUCTION

In nanometer technologies, variability is becoming one of the leading causes for chip failures and delayed schedules [1]. In general, the root causes can be (1) Process variability (e.g., Vth variations, thickness variations [2], gate L (Length) and W (Width) variations, etc.), (2) Operational variability (multiple design modes, Dynamic Voltage and Frequency Scaling [3], power gating, etc.), (3) Operation-dependent variability (e.g., PVT (Process, power-supply Voltage, and Temperature) variations [4], crosstalk [5], IR-drop [6], etc.) and (4) Long-term variability (aging [7] effects, namely due to NBTI (Negative Bias Thermal Instability) [8], radiation, etc.). Variability causes increasing uncertainty in system behavior, namely on its performance.

Moreover, variability also decreases circuit dependability, i.e., its ability to deliver the correct functionality within the specified time frame. Hence, lower circuit’s dependability and reliability [9][10] are to be expected, when moving to low nanometer range. Clearly, PVT (PVT and Aging) variations degrade circuit’s dependability. Conservative approaches, with wider relative circuits’ time slacks, are required when technology is down scaled [11]. Time slack (τ slack) is the time period exceeding the longest propagation delay time between memory elements, to accommodate PVT variations.

Semiconductor aging [7] causes long-term performance degradation, and may activate physical defects latent in production. For instance, MOS threshold voltage (Vth) parametric variations may be induced either due to holes trapped in the thin gate oxide (SiO2), or due to Si/SiO2 surface change. The first effect tends to decrease VthN and to increase |VthP| (NMOS and PMOS Vth, respectively). The second effect, under high radiation, tends to increase VthN and |VthP|. Hence, Vth shifts will modify the timing performance of digital circuits.

NBTI has been identified as the dominant long-term effect in nanometer CMOS technologies [8][12]. It primarily affects PMOS transistors, increasing |VthP| along the time. This will ultimately cause a delay fault. For safety-critical (e.g., automotive) and mission-critical applications (e.g., space), this must not occur. Performance failure prediction must be implemented.

The purpose of this paper is to present an adaptive error prediction flip-flop (AEP-FF), to be used in a performance failure prediction methodology for safety-critical, high-performance systems. This methodology is based on monitoring long-term performance degradation of CMOS digital systems. In particular, key flip-flops (FF) are chosen to include the aging sensor functionality and monitor FF data input late transitions. The proposed AEP-FF performs locally all the monitoring procedure, reducing to a minimum the global interconnections for aging sensor insertion. The observation (or guard-band) interval, tOb, at the end of the clock cycle, is defined by design, so the time period in which abnormal delays are observable is user’s defined. However, with the novel sensor’s architecture, its sensitivity (measured...
by $t_g$ increases with its PVT variations. This way, the sensor FF will adapt and increase the guard-band, as circuit variability increases with aging.

The paper is organized as follows. Section II briefly reviews previous work on aging and its impact on system performance, and on aging sensors. In section III, the proposed flip-flop architecture is presented. Section IV describes the aging sensor methodology, including sensor insertion procedure and circuit reconfiguration. In section V, simulation results are presented. Finally, conclusions are summarized in section VI.

II. PREVIOUS WORK

The impact of NBTI on digital circuit performance has been under a significant research effort [8][12][13]. NBTI modeling in static and dynamic operation has been proposed [13]. Solutions to enhance tolerance to $V_{dd}$ degradation have also been presented [7][14], and [12] shows how design optimization can restrict NBTI performance degradation. Moreover, various aging sensor topologies have been presented [15][16][17][18] to globally detect circuit’s performance degradation.

A different aging sensor approach is the circuit failure prediction technique proposed by M. Agarwal et al. in [19]. The underlying idea is to anticipate system failure, before it really occurs. Their major application was to reduce the pessimistic worst-case delay to accommodate PVT variations, which significantly limits system performance.

More recently, an aging sensor methodology focusing a different application was presented in [11][20][21][22]. Here, aging monitoring is performed during product lifetime under heterogeneous VT (power supply Voltage and Temperature) variations. The methodology includes (a) programmable aging sensor design, resilient to aging and showing low sensitivity to PVT variations, and (b) a monitoring procedure, with automatic sensor insertion.

Nevertheless, the proposed methodology has limitations, namely (1) sensor must have better performance and lower sensitivity to PVT variations than the CUT (Circuit Under Test); (2) methodology is based in monitoring time interval (guard-band) that needs to be synchronous with the clock; (3) sensors are active in pre-defined periods during which the CUT critical paths may not be activated. The present paper presents a novel FF architecture with built-in aging sensor, based on the previous work done in [21]. However, a different approach is used, in order to deal with the reported limitations of previous aging sensors topologies.

III. ADAPTIVE ERROR-PREDICTION FLIP-FLOP

In a synchronous digital system, data signals processed in combinational modules are registered in storage elements (FF). However, if a signal propagation delay is large enough (e.g., due to PVTA variations), it will induce a de-synchronization effect due to the increased difference between the critical path propagation delay, $T_{pd}$, and the clock distribution network delay, $\tau$. Consequently, an aging sensor must monitor performance degradation locally, at key memory cells, where synchronization errors start to occur.

We refer these FF as CME (Critical Memory Elements). Aging sensors must be integrated in FF terminating critical or near-critical paths (CP). The sensor monitors late transitions at the FF’s data input.

The topology and principles of operation of the proposed Adaptive Error-Prediction Flip-flop (AEP-FF) are shown in Figures 1 and 2, respectively. The Delay Element (DE) delays data signals captured at the Master Latch output, during CLK low state. The Stability Checker (SC) analyzes data transitions during CLK high state. This way, the DE propagation delay is the effective observation (or guard-band) interval, $t_g$, used by the sensor. Late transitions at FF data input (propagated to the Master Latch output) will be identified by the SC. As it will be shown later, SC has on-state retention logic, to discard the use of an additional latch to store the aging sensor output signal (AS_OUT).

Using the flip-flop data signal at the output of the Master Latch to drive the DE, instead of the FF input data signal, D, as in previous aging sensor architectures [19][20], simplifies DE design. Basically, the new delay element is a simple buffer that introduces a delay to create a virtual guard-band where late transitions at FF data input are signalized. The loading effect of the sensor is inside the FF; hence, it does not explicitly impact the signal path. Moreover, if PVTA variations occur in DE, this virtual guard-band will increase accordingly. It is possible to create a power-on state for the DE, and activate it in short periods to maintain its propagation delay approximately constant when compared with the circuit’s performance degradation. Even when the sensing operation is always ON, the increased workload of the DE will cause sensor’s guard-band to increase as aging effects cumulatively degrade DE performance. This way, the sensor’s sensitivity is adapted with the cumulative aging.
degradation of the circuit. Another advantage is that the guard-band signal does not need to be distributed as a second balanced clock to the sensing FFs, as in [19][20].

Figure 3 presents three typical DE architectures. These buffers have different delay capability and different aging performance degradations. The DE architecture should be chosen according to the technology used: e.g., architecture (a) does not work in nanometer technologies, (b) and (c) do; the time slack margin, $t_{\text{slack}}/T_{\text{clk}}$, is also relevant. In fact, for high performance circuits (low $t_{\text{slack}}/T_{\text{clk}}$), architecture (c) is preferred. Moreover, changing W/L transistors ratios also change the sensor’s effective guard-band, $t_g$. Unlike in [20][21][22], $t_g$ is not programmable, it is defined at design time. However, $t_g$ is adaptive with PVTA variations, enhancing sensor’s detection sensitivity.

The novel Stability Checker (Figure 4) is implemented with dynamic CMOS logic and has built-in on-retention logic. During CLK low state, and considering that AS_out signal is low, X and Y nodes are pulled up (making AS_out signal to stay low). When CLK signal changes to high state, M3 and M4 are OFF, and according to Delayed_DATA signal, one of the nodes X or Y changes to low. If, during the high state of the CLK, a transition in Delayed_DATA occurs, the high X or Y node is pulled down by transistor M2 or M5, respectively, driving AS_out to go high. From now on, M9 transistor is OFF. Hence, X and Y nodes are not pulled up during CLK low state, unless the active low RESET signal is active. X and Y nodes remain low, helped by transistors’ M3 and M4 activation during AS_out high state. For the RESET signal to restore the cell’s sensing capability, it must be active, at least during the low state of one clock period.

The proposed SC architecture, with the on-retention logic implemented with transistors M3, M4, M8 and M9, does not need an additional latch to retain the SC output signal when it’s active. However, additional research must be pursued to analyze the SC operation in the presence of cross-talk noise, especially when CLK signal is activated, as false-positive errors may be signalized. Moreover, only FF’s internal clock signal is triggering the beginning of the observation interval, $t_g$. As mentioned, guard-band interval is the DE propagation delay, ultimately limited by half the clock cycle (when CLK signal is high). We refer it as a virtual guard-band because there is no signal explicitly representing the observation interval. Each sensing FF will have its own unique PVTA-dependent guard-band (each local DE may age differently).

![Figure 3. Delay element typical architectures.](image)

![Figure 4. Stability checker architecture with on-retention logic.](image)

**IV. AGING MONITORING PROCEDURE**

The monitoring procedure comprises (1) aging sensors (AS) characteristics, (2) overhead attributes, (3) monitoring effectiveness, and (4) sensor insertion criteria, algorithm and tool. Valued AS characteristics are: low power consumption, area overhead and performance degradation, and sensitivity to PVTA variations (increasing $t_g$ as variability increases). Due to DE input location (at the Master Latch output), the performance degradation of the AEP-FF is negligible.

Overhead attributes are hardware overhead. In terms of hardware, one metric is defined: global overhead, $GO = n_{AS}/m$, where $n_{AS}$ is the number of inserted AS and $m$ is the total number of memory elements in the CUT.

In terms of monitoring effectiveness, two issues are critical: looking at the right place, and long enough to uncover abnormal delays (the activation of the critical paths is an important task, but is out of the scope of this work). Hence, sensors must be inserted at the end of critical (CP) and near-critical delay paths. How many near-critical paths should be sensed? A CP delay variation margin ($\alpha$) is user’s defined, in order to allow the decision on how many and which near-critical paths should be sensed.

The following sensor insertion criteria are considered:
1. CP identification, and signal path delay map building (using a Statistical Timing Analysis tool, e.g., PrimeTime™). The path with the longest propagation delay time ($t_{\text{delay}}$) is identified, and all signal paths are ranked, according to their delay times.
2. CME identification, according to user’s defined ($\alpha$) monitoring effectiveness. Selected CME end all signal
paths whose propagation delay exceeds [α, tpdmux]. Typically, α=90%, to consider PVT variations.

3. Circuit reconfiguration at logic-level, to automatically replace critical flip-flops (CME) with AEP-FFs.

An algorithm to apply the three criteria and perform sensor insertion is embedded in a proprietary tool, DyDA (Dynamic Delay Analyser) [4]. DyDA (1) collects Prime Time™’s output data, (2) generates path delay distribution and potential CME insertion, and (3) performs netlist reconfiguration (upon user’s decision (α)), replacing few FF by AEP-FFs. To help user’s decision on defining α, DyDA tool generates a ranked list of all CMEs, based on their input cone’s critical path and a graph with the distribution of the memory elements’ input cone CP and all paths distribution.

Figure 5 depicts path delay and CME distribution curves for a 2-stage, 4-bit Pipeline Multiplier (PM) (AMS 0.35µm), used as CUT1. PM is a 36 FF, 52 gate structure, 9 (8 data, 1 CLK) inputs, 8 outputs (Fig. 6). As shown, for the longest paths, with α=90%, 5 CME are needed. If a larger safety margin is required (α=90%), 4 additional CME are needed. Note that more than one large delay path converge to the same FF. In fact, in this case, 23 long signal paths converge in 5 CME of the 20 FF at the end of pipeline stage 1.

V. SIMULATION RESULTS

Simulation results are presented for CUT1 (PM) (0.35µm and PTM 65nm [23]) and CUT2 (inverter chain creating a CP example) (Fig. 7) (PTM 65nm). A 1V nominal Vdd is considered, with variations from 1.2V to 0.8V. T variations are restricted from 27°C to 150°C. WC (worst case) VT conditions are Vdd=0.8V, T=150°C. Aging fault injection, by Vap modification, is performed by Spice VTHO parameter modulation. NBTI aging is assumed to equally degrade Vap of all PMOS transistors.

Figure 8. CP’s propagation delay variations (ΔtLH and Δtmux) in CUT2 with aging (AVap), for nominal (NC) and worst-case (WC) conditions.

Figure 7. CUT2 (40 inverters’ chain, creating a typical CP example).

Considering 65nm PTM CUT2 simulations with HSPICE, the propagation delay time of CUT2 CP at nominal conditions (NC) leads to 326ps (L-H transition). WC VT conditions lead to tLH=805ps, which represents a serious degradation, as compared to nominal tLH. The clock period is defined as Tclk=τLH+τclock, where τLH=τSU+τSU0, and τSU is the FF set-up time, namely τSU=39ps. Hence, Tclk=844ps. In order to accommodate process variations, a time slack was set as 30% of τLH. Therefore, Tclk = 1.1ns will guarantee correct operation under worst case PVT conditions.

Simulations were carried out to compute propagation delay degradation due to NBTI-induced aging. The impact of up to 30% variations in [Vap] of the PMOS in the CUT on % variations of tLH and tHL was evaluated (Fig. 8): 30% Vap variation of leads to ΔtLH=16% at VT NC. Worst case VT scenario leads to higher delay degradation (35%).

For CUT2, AS insertion was carried out, with selective AEP-FF insertion, implemented with the DE shown in fig. 3-c) (with minimum-size PMOS transistors). The AEP FF performance degradation was evaluated: only 5% set-up time degradation was measured (τSU_AEP_FF = 41ps) for WC conditions (when compared to Tclk, it’s irrelevant), while no hold-time degradation was measured (τSU0_AEP_FF = 9ps).

Error-prediction by the AEP-FF starts to occur for 10% degradation in Vap and for 9% degradation in the critical path propagation delay (L-H transition). The sensor continues its error-prediction until the error occurs (when τclock is violated). The sensor’s effective virtual guard-band, τg, was computed under VT variations (Fig. 9) (only L-H transition is shown, as H-L transitions’ results are similar) and under Vap aging variations (Fig. 10). Moreover, when analyzing τp variation with VTA variations (Figs. 9 and 10), unlike in [19][20], our proposed sensor increases the guard-band interval as variability in the circuit increases with V reduction and T and [Vap] increase. This is an important result, because Vap aging degradation, or V depletion, or T increase, unlike other aging sensor solutions, work in favor of the error-prediction methodology. In fact, the AS becomes

Figure 8. CP’s propagation delay variations (ΔtLH and Δtmux) in CUT2 with aging (AVap), for nominal (NC) and worst-case (WC) conditions.
more sensitive to CUT delay variations, as VTA variability increases. Previous works [19][20] put a heavy burden on AS design, to guarantee that the AS sensitivity to VTA variations does not reduce significantly the detectability window of the sensor.

![Graph](image1)

Figure 9. AS effective guard-band variation (L-H transition). (a) under temperature variations; (b) under power-supply voltage variations.

Furthermore, in [19][20][21] SC operation significantly reduces the effective t\textsubscript{g} of the sensor. In high performance circuits, the SC must also react fast to maintain the predictive error detection capability. In the proposed AEP-FF, the data signal that drives the SC is always stable, due to the fact that Delayed\_DATA signal is analyzed at the output of the master latch of the FF. Hence, there is no limitation imposed by the eventually slow operation of the SC.

For the CUT1 Pipeline Multiplier (PM, Fig. 6), implemented with 65nm PTM, the 5 most critical memory elements are the same as the ones identified by DyDA on the 0.35\textmu m version (Fig. 5). Fig. 11 depicts their WC propagation delay variation with NBTI-induced aging.

![Graph](image2)

Figure 11. Propagation delay degradation of 5 CP ending at 5 CME in CUT1, under 5% variations in the PMOS V\textsubscript{thp}, under worst case VT conditions.

After PM circuit reconfiguration, replacing the 5 CME with the AEP-FF, HSPICE simulations were performed with T\textsubscript{clk}=780ps, where 60% corresponds to the propagation delay of the CP under WC conditions, with T\textsubscript{slack}=40% T\textsubscript{clk} to accommodate WC PVT variations. Fig. 12 shows, for PM (CUT1) and WC VT conditions, the abnormal delays detection ranges at the inputs of the 5 CME (slowest HL or LH transitions), showing how effective the aging sensor is. The sensor can detect from 16% to 40% V\textsubscript{thp} degradation.

![Graph](image3)

Figure 12. Effective detection ranges in percent variations of V\textsubscript{thp} degradation variations, for CUT1 under NC VT conditions at the 5 CME

For safety-critical applications, the sensor must provide reliable operation also in the presence of P (Process) variations. Monte Carlo (MC) simulations under WC VT conditions were carried out. In all MC simulations, a Gaussian distribution with +/-3\textsigma variation of +/-10% of the nominal values is assumed [24] for 3 MOSFET parameters: L\textsubscript{eff}, tox, and V\textsubscript{th}. MC simulations were run varying MOSFET parameters in the CUT and in the sensor’s transistors. For each set of MC simulations, 30 runs were performed. Under P variations, detection on fast samples requires high V\textsubscript{thp} degradation. We define the Detection Probability (DP) as the percentage of the 30 MC runs that detect the abnormal delay. Fig. 13 presents MC results for DP in the 5 CME of CUT1.

The monitoring methodology was used to define sensor insertion in a set of benchmark circuits (Table 1): ITC’99 benchmarks, 2-stage 4-bit Pipeline Multiplier (PM 4-2), 4-stage, 16-bit PM (PM 16-4) and a simple PIC controller. Results show that few AEP-FF need to be inserted. Of course, for sequential circuits it is less rewarding. However, propagation delay distribution depends on circuit synthesis,
which can be used to restrict the global overhead (GO), making less circuit nodes prone to unsafe time response.

![Monte Carlo Results for Detection Probability (DP)](image)

**Figure 13.** Monte Carlo statistic results for first detections at the 5 CME for percent variations of $V_{thp}$ degradation variations, in CUT1 under WC VT conditions.

**TABLE I. AGING SENSOR INSERTION USING DYDA TOOL (AMS 0.35µM)**

<table>
<thead>
<tr>
<th>Circuit</th>
<th># FF</th>
<th># AEP-FF</th>
<th>GO (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD1</td>
<td>5</td>
<td>2</td>
<td>40.0</td>
</tr>
<tr>
<td>BD3</td>
<td>30</td>
<td>22</td>
<td>73.3</td>
</tr>
<tr>
<td>BD6</td>
<td>9</td>
<td>3</td>
<td>33.3</td>
</tr>
<tr>
<td>BD9</td>
<td>45</td>
<td>23</td>
<td>51.1</td>
</tr>
<tr>
<td>Bat. PM 4-2</td>
<td>36</td>
<td>5</td>
<td>13.9</td>
</tr>
<tr>
<td>Simple PIC</td>
<td>66</td>
<td>16</td>
<td>24.2</td>
</tr>
<tr>
<td>Bat. PM 16-4</td>
<td>480</td>
<td>7</td>
<td>1.5</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

In this paper, a new Adaptive Error-Prediction Flip-flop (AEP-FF) was presented, to be used in a performance failure prediction methodology for safety-critical applications. It was shown that sensor’s guard-band increases with $V_{thp}$ depletion, T increase and NBTI-induced aging. The AEP-FF includes new DE and a new SC architecture with on-retention logic feature. AEP-FF architecture reduces the performance degradation overhead due to sensor insertion, as compared to previous solutions. Moreover, SC operation does not limit the effective detection window of the sensor, as in previous AS topologies. As sensors age, sensor’s performance degradation works in favor of the predictive error detection. The methodology includes aging sensor design and insertion, using DyDA tool. Simulation results demonstrate AEP-FF robustness on detecting late transients at FF input data signal, even in the presence of its own aging.

Future work includes cell library design, for complete methodology automatic insertion, workload-dependent NBTI effects, analysis of sensor’s robustness in the presence of cross-talk noise and more extensive MC analysis to account for process variations, and will be reported in the future.

ACKNOWLEDGMENT

The work has been partially supported by ENIAC SE2A Project, and by FCT (INESC-ID multianual funding) through the PIDDAC Program.

REFERENCES


