Abstract—Circuit aging refers to the deterioration of circuit performance over time. All portions of a System-on-Chip (SoC), analog, digital logic, and memory, are affected by aging. In this survey, three main aging mechanisms will be introduced: Bias Temperature Instability (BTI), Time Dependent Dielectric Breakdown (TDDB), Hot Carrier Injection (HCI). This survey also introduces a concept of circuit failure prediction for a dominant PMOS aging mechanism induced by Negative Bias Temperature Instability (NBTI). NBTI-induced PMOS aging slows down PMOS transistors over time. As a result, the speed of a chip can significantly degrade over time and can result in delay faults. Finally we give some Aging management methods and comments.

Keywords: BTI; TDDB; HCI; Circuit Failure Prediction

I. INTRODUCTION

Three main aging mechanisms are commonly accepted: Bias Temperature Instability (BTI), Time Dependent Dielectric Breakdown (TDDB), and Hot Carrier Injection (HCI).

Bias temperature instability (BTI) is a phenomenon that causes threshold voltage shifts over long periods of time, eventually causing the circuit to fail to meet its specifications. The word “bias” refers to the fact that this degradation is heightened by the application of a bias on the gate node of a transistor.

Time-dependent dielectric breakdown (TDDB) in the gate oxide is a reliability phenomenon in gate oxides that results in a sudden discontinuous increase in the conductance of the gate oxide at the point of breakdown, as a result of which the current through the gate insulator increases significantly. This phenomenon, illustrated in the figure below, is of particular concern as gate oxide thicknesses become thinner with technology scaling, and gates become more susceptible to breakdown.

Degradation due to hot carrier injection (HCI) appears mainly in NMOS devices when a large drain-to-source voltage and gate-to-source voltage is applied. HCI manifests itself as an increase in threshold voltage and a decline in channel mobility, leading to degradation in transistor drain current. HCI is caused by various effects: the traditional explanation was based on impact ionization, but more complex effects are seen in nanometer-scale technologies. With supply voltages leveling off even as geometries shrink HCI will worsen in future technologies, and is likely to be the dominant effect for long-term failures.

The paper we read introduces an ideal circuit failure prediction (CFP) to solve BTI problems.

The idea behind circuit failure prediction is to predict the occurrence of a failure during normal system operation before the appearance of any error that can result in corrupt system data and states. This is in contrast to classical error detection where a failure is detected after errors appear in system data and states. Failure prediction can be performed in multiple ways - the basic principle is to insert a wide variety of “sensors” at various locations inside a chip. During normal system operation, these sensors collect information about various system parameters over time. Examples of such system parameters range from temperature, voltage, ring oscillator delays to complex relative timing relationships among logic signals. The data collected by the sensors is analyzed on-chip or off-chip to identify “anomalies” and predict failures. Sensor designs and data analysis techniques can widely vary depending on failure sources and their sensitivities. The purpose of this paper is to introduce the concept of circuit failure prediction and apply this concept to a dominant reliability mechanism - PMOS transistor aging induced by Negative Bias Temperature Instability (NBTI).

II. BIAS TEMPERATURE INSTABILITY (BTI)

Bias temperature instability is a phenomenon that causes threshold voltage shifts over long periods of time, eventually causing the circuit to fail to meet its specifications.

The phenomenon of negative bias temperature instability (NBTI) can be illustrated with the help of a simple circuit, an inverter, illustrated in Fig.2.1 (a). When a PMOS transistor is biased in inversion (for example, when the input of the inverter is at logic 0), interface traps are generated due to the dissociation of Si - H bonds along the substrate-oxide interface, as illustrated in Fig.2.1 (b). The connection of this mechanism to thermal effects is that the rate of generation of these traps is accelerated by elevated temperatures, and therefore, increased on-chip temperatures can directly affect the lifetime of a chip. The time for which the transistor is stressed is another factor that increases the level of degradation. These traps cause an increase in the threshold voltage, and a reduction in the saturation current of the PMOS...
transistors. This effect, known as NBTI, has become a significant reliability issue in high-performance digital IC design, especially in sub-130 nm technologies. An increase in causes the circuit delay to degrade, and when this degradation exceeds a certain amount, the circuit may fail to meet its timing specifications. The rate constants of the reactions that define NBTI are dependent on temperature, and are worsened at elevated temperatures.

Fig 2.1

A corresponding and dual effect, known as positive bias temperature instability (PBTI) can be seen for NMOS devices, for example, when the input to an inverter is at logic 1, and a positive bias stress is applied across the gate oxide of the NMOS device. Although the impact of a stressing bias on PBTI is lower than NB, PBTI is becoming increasingly important in its own right. Moreover, techniques are developed to reduce NB can contribute to increasing PBTI. For example, if the NMOS transistor is at logic 1, corresponding to the signal at its gate node. When the stress is removed, the threshold voltage is seen to recover towards its original value. Analytical models for the change in threshold voltage are provided. For NBTI, the degradation in the threshold voltage is proportional to the probability p that the signal at the gate of the transistor is at logic 0, corresponding to the proportion of time that the transistor is likely to be under stress. A similar model for PBTI concludes, analogously, that the degradation is proportional to 1-p.

Fig 2.2

Static methods incur high overheads due to the padded delays: instead, dynamic run-time methods may be used with a small static overhead. Moreover, it can be observed that over the life of the circuit, the delay increases but the leakage power decreases. As a result of the latter, the circuit may be well within its power budget late in its life, implying that one must overdesign the resources used to control power and temperature based on the requirements at the beginning of life.

The approach in attempts to dynamically adjust the circuit delay using a combination of adaptive body biases and adaptive supply voltages. This method begins with a smaller initial delay padding than the static method, and alters the body biases and supply voltages over time using a sensor-based method, ensuring that the timing specification is met, while staying close to the power budget throughout the circuit lifetime. In this case, a time sensor is used, but silicon odometer methods could also be adapted to similar schemes.

Fig 2.2 below shows the change in the delay and power as a function of time for five scenarios. The nominal case, which meets the delay specifications, violates the delay constraint almost immediately and is not viable. The static padding case uses delay padding to meet the timing constraints at the end of life, and incurs significant power overheads, even accounting for the fact that the leakage reduces with time. The adaptive approach uses adaptive body biases and adaptive values; the slight initial padding is due to the discreteness of the cell library. The hybrid approach, which is the best solution, uses a mix of initial delay padding and adaptive behavior. The sum of dynamic and leakage power is seen to show the best overall trend.
III. **Time Dependent Dielectric Breakdown (TDDB)**

Time-dependent dielectric breakdown (TDDB) in the gate oxide is a reliability phenomenon in gate oxides that results in a sudden discontinuous increase in the conductance of the gate oxide at the point of breakdown, as a result of which the current through the gate insulator increases significantly. This phenomenon, illustrated in the figure below, is of particular concern as gate oxide thickness becomes thinner with technology scaling, and gates become more susceptible to breakdown.

Gate oxide breakdown leads to alterations in the DC noise margins for read, write, and retention, as well as in the read and write access times.

![Figure 3.1](image)

**IV. Hot Carrier Injection (HCI)**

Degradation due to hot carrier injection (HCI) appears mainly in NMOS devices when a large drain-to-source voltage and gate-to-source voltage is applied. HCI manifests itself as an increase in threshold voltage and a decline in channel mobility, leading to degradation in transistor drain current. HCI is caused by various effects: the traditional explanation was based on impact ionization, but more complex effects are seen in nanometer-scale technologies. With supply voltages leveling off even as geometries shrink, HCI will worsen in future technologies, and is likely to be the dominant effect for long-term failures.

**V. Circuit Failure Prediction**

**A. Definition**

Circuit Failure Prediction is to predict the occurrence of a failure during normal system operation before the appearance of any error that can result in corrupt system data and states. Normally it can be performed in multiple ways – the basic principle is to insert a wide variety of “sensors” at various locations inside a chip. During normal system operation, these sensors collect information about various system parameters over time. Examples of such system parameters range from temperature, voltage, ring oscillator delays to complex relative timing relationships among logic signals. And then the data is analyzed on-chip or off-chip to identify “anomalies” and predict failures.

**B. Comparison between error detection**

And this method is really different from our normal error detection method and the differences are shown in table 4.1.

The major differences are:

1. Circuit Failure Prediction mainly predicts error so it’s usually handled before error appears while normal error detection can only detect the error after it really happens.
2. Failure prediction enables a system to initiate corrective measures before system data and states actually get corrupted by errors – this prevents error propagation and error detection latency problems associated with classical error detection.
3. The biggest limitation is that failure prediction cannot predict all failures, e.g., radiation-induced soft errors. However, it is suitable for failures caused by major reliability mechanisms such as NBTI-induced transistor aging because of the gradual nature of degradation. And in this part we mainly talk about the NBTI effects.
4. It usually takes less time as we only take data during some guardband interval (say 15days) while it may take really long time for error detection to detect the error. And thus it costs less and more efficient.

<table>
<thead>
<tr>
<th>Circuit failure prediction</th>
<th>Error detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before errors appear</td>
<td>After errors appear</td>
</tr>
<tr>
<td>No corrupt data or states</td>
<td>Corrupt data &amp; states</td>
</tr>
<tr>
<td>Can collect data over several cycles before predicting failure</td>
<td>Long error detection latency causes data integrity problems.</td>
</tr>
<tr>
<td>Generally inexpensive</td>
<td>Generally expensive</td>
</tr>
<tr>
<td>Self-diagnoses possible</td>
<td>Limited self-diagnoses</td>
</tr>
<tr>
<td>Incorrect prediction can be a problem</td>
<td>Insufficient coverage can be a problem</td>
</tr>
<tr>
<td>All failures not predictable</td>
<td>General applicability</td>
</tr>
<tr>
<td>Both can be efficiently combined</td>
<td></td>
</tr>
</tbody>
</table>

**C. Basic steps for CFP**

The basic step for overcoming NBTI-induced PMOS aging is shown in the figure 4.1. Firstly the system starts off with a small worst-case timing guardband that guarantees correct circuit operation even with worst aging over a short period of time, e.g., 15 days. We refer to this timing guardband as the guardband interval. The choice of this period is dependent on overall circuit properties and will be discussed later.

During this 15-day period of operation, various parts of a chip will age at various rates depending on the workload, voltage and temperature profiles. During this time, specially
designed on-chip sensors collect lots of data about relative aging of various circuit paths.

At the end of 15 days, the data collected by the sensors is analyzed to check whether there has been enough aging that requires adjustment of the 15-day timing guardband. If the aging is enough, we use a self-adapt or self-correct system to adjust the guardband.

Figure 5.1 Circuit Failure Prediction for NBTI-induced aging

D. Sensor Design

One of the major problems in the CFP is to design the sensor and normally there are two ways to design the embedded predicting sensors.

Our sensor design approach is to modify a standard flip-flop by inserting a "monitoring" circuit block which detects any ‘significant’ shifts in delay of the combinational logic whose output is connected to the data input of that flip-flop. The monitoring circuit block is based on the concept of stability checking during the guardband interval by detecting signal transitions during the guardband interval, also referred to as guardband violation. The guardband violation means one or more paths in the combinational logic have age enough to creep into the guardband interval and now is very close to turning into a delay fault.

Figure 5.2 Flip-flop with built-in circuit failure sensor

E. The stability checking based sensor

An aging sensor integrated inside a flip-flop shown in Figure 4.3 mainly has three components:
1. A stability checker
2. A delay element to create an interval during which stability checking is performed
3. An output latch to store stability checking results

The delay element and the output latch may be shared among multiple flip-flops for power and area savings.

Figure 5.3 Block Diagram for Stability checking based aging sensor.

And the logic level design for these parts is as follows:

(1) Stability Checker

Figure 4.4 shows the design of a stability checker. In the beginning of a clock cycle when Clock = 1 (i.e., Clock_b = 0), PMOS transistors T1 and T5 are on (NMOS transistors T3 and T7 are off), and the stability checker output Out = 0. This is called the precharge phase of the stability checker.

The delay element (Delay) in Fig.4.4 introduces a delay of ‘Telk/2 – Tg’ (assuming 50% duty cycle of Clock). This guarantees that transistors T3 and T4 (and transistors T7 and T8) are both on during the guardband interval Tg (called the evaluate phase).
During the guardband interval, PMOS transistors T1 and T5 are turned off. Stability Checker output Out becomes 1 if and only if the combinational logic output OUT transitions from 1 to 0 or 0 to 1 once or multiple times during the guardband interval. And the latch in Figure 5.3 is to hold these changes and will analyze later.

This technique is simple and the power impact of the delay element here can be smaller than the delay required in Fig. 5.5 since the activity factor of the combinational logic output is lower than the clock. However, if the aging sensors are mostly turned off and are turned on as and when required, then the power impact will be very small for design shown in Fig. 5.5 design as well. The design in Fig. 5.6 is prone to invalidation due to hazards (similar to invalidation during path delay testing). Moreover, unlike the design in Fig 5.3, the delay element cannot be shared among multiple stability checkers and contributes towards larger area and power. For both sensor designs, additional resolution can be obtained by adding additional delay elements.

(2) Delay Element Design

The Delay element part design is shown in figure 5.5.

Since NBTI-induced PMOS aging is a slow process, aging sensors can be turned off most of the time to ensure that they don’t age significantly. This requires an additional slow global input signal, Monitor. (shown in the figure below).

When Monitor = 1, aging monitoring is turned on and the delay element outputs a delayed version of Clock_b. When Monitor = 0, aging monitoring is turned off and the delay element produces 1. Turning aging monitoring off also reduces power consumption.

F. Pre-sampling based sensor

An alternative aging sensor design is based on pre-sampled outputs shown in Fig 5.6. The idea is to pre-sample the combinational logic output (using a delay element) and compare it with the actual captured value using an XOR gate. If the guardband is violated due to aging, then the two samples will be different and a mismatch will be reported by the XOR.

This technique is simple and the power impact of the delay element here can be smaller than the delay required in Fig. 5.5 since the activity factor of the combinational logic output is lower than the clock. However, if the aging sensors are mostly turned off and are turned on as and when required, then the power impact will be very small for design shown in Fig. 5.5 design as well. The design in Fig. 5.6 is prone to invalidation due to hazards (similar to invalidation during path delay testing). Moreover, unlike the design in Fig 5.3, the delay element cannot be shared among multiple stability checkers and contributes towards larger area and power. For both sensor designs, additional resolution can be obtained by adding additional delay elements.
Table 5.2 Flip-flop-level performance, power and area tradeoffs of aging sensors (chip-level penalties are lower).

<table>
<thead>
<tr>
<th>Tradeoffs/ Penalties</th>
<th>No sharing of Delay element &amp; latch</th>
<th>Delay element &amp; latch shared among 4 blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance penalty</td>
<td>&lt; 1%</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Power penalty when aging sensors are off (most of the time)</td>
<td>0.3%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Power penalty when aging sensors are on (small fraction of time)</td>
<td>12.5%</td>
<td>7.5%</td>
</tr>
<tr>
<td>Area (Transistor count per sensor)</td>
<td>51</td>
<td>24</td>
</tr>
<tr>
<td>Global signal</td>
<td>1 (slow)</td>
<td>1 (slow)</td>
</tr>
<tr>
<td>Minimum guardband Tg</td>
<td>20ps</td>
<td>20ps</td>
</tr>
</tbody>
</table>

Table 5.3 below summarizes chip-level costs associated with the aging sensors. Overall chip-level power penalty is very small because the sensors do not need to be used for all flip-flops and do not need to be activated all the time.

And the example is a synthesized OpenRISC processor design (available from www.opencores.org) shows that roughly 40% of all flip-flops require built-in sensors for pessimistic situations in which circuit path delays can increase by up to 25% due to transistor aging. This estimate is made by identifying each flip-flop with D-input connected to at least one circuit path with slack less than 25% of the total cycle time. In that case, the chip-level power penalty of using aging sensors will be 4% when all aging sensors are on. Since, aging sensors are expected to be on for only a small fraction of time, chip-level average power penalty further reduces to 0.4% if the aging sensors are on 10% of the time (which is a pessimistic assumption). The flip-flop with built-in aging sensor can be a standard cell in the technology library. Hence, we do not expect any major impact on design and test flows.

2) Extended Guardband for Reliable Aging Sensor Operation

If the combinational logic output transitions very close to either the rising or the falling edge of the guardband interval formed by the $\text{Clock}_b$ and delayed $\text{Clock}_b$ signals, then a problem will rise up the transition may not be reliably detected by the stability checker. And this is normally treated as setup time violation of the stability checker block. To overcome this problem, we add a small margin to the aging guardband interval to create an extended guardband. And through our simulation, we can see the extended guardband is about 20ps for the design and it will be shown later.

![Extended guardband](image)

Figure 5.7 Additional 20ps setup time margin to reliably detect signal transitions on combinational logic output.

3) Guardband Reduction Effectiveness using Aging Sensors

Table 5.4 presents analytical results for the amount of guardband reduction that can be obtained by using our failure prediction technique for systems with clock frequencies of 1 GHz, 2 GHz and 3 GHz. Here we sweep the initial guardband from 1 day to 60 days. In addition, we sweep the amount of 7-year delay degradation due to transistor aging from 5% to 40% of the clock period. Here is a sample calculation to illustrate how the numbers in Table 5.4 are obtained.

We have the formula that

$$\frac{\Delta \tau}{\tau} = K t^{0.25}$$

Where $\tau$ is the gate delay, $\Delta \tau$ is the change in gate delay due to aging, $t$ is the time in days and $K$ is a constant proportionality.

![Extended guardband](image)

Table 5.3 Chip-level performance, power and area tradeoffs of aging sensors for synthesized OpenRISC core

<table>
<thead>
<tr>
<th>Speed penalty</th>
<th>&lt; 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>% flip-flops with built-in aging sensors (pessimistically assuming 25% increase in path delays due to aging)</td>
<td>&lt; 40%</td>
</tr>
<tr>
<td>Average power penalty (aging sensors on 10% of the time)</td>
<td>&lt; 0.4%</td>
</tr>
<tr>
<td>Additional global signal</td>
<td>1 (slow)</td>
</tr>
<tr>
<td>Scan path reuse</td>
<td>Possible</td>
</tr>
<tr>
<td>Die size increase</td>
<td>Not expected</td>
</tr>
<tr>
<td>Design flow impact</td>
<td>Minimal</td>
</tr>
</tbody>
</table>
Table 5.4 Percentage reduction in guardband enabled by circuit failure prediction using the aging sensor

<table>
<thead>
<tr>
<th>Clock period</th>
<th>7 yr guardband (% of clock period)</th>
<th>Initial guardband interval (days)</th>
<th>1</th>
<th>5</th>
<th>15</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHz</td>
<td></td>
<td></td>
<td>5</td>
<td>46</td>
<td>59</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>66</td>
<td>52</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>78</td>
<td>64</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>81</td>
<td>67</td>
<td>56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHz</td>
<td></td>
<td></td>
<td>5</td>
<td>6</td>
<td>-1</td>
<td>-9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>66</td>
<td>52</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>78</td>
<td>64</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>81</td>
<td>67</td>
<td>56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHz</td>
<td></td>
<td></td>
<td>5</td>
<td>6</td>
<td>-1</td>
<td>-9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>66</td>
<td>52</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>78</td>
<td>64</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>81</td>
<td>67</td>
<td>56</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We will show how these values are calculated in the table. For example, considering a system with 1GHz and suppose that aging over 7 years imposes a guardband of 25% of the clock period (= 250ps). Suppose that we use circuit failure prediction instead, using our aging sensor for a system, starting off with a 60-day guardband (instead of a 7-year guardband).

7 year reduction is about 250ps
Degradation during first 60 days is:

\[ 0.25 \times \left( \frac{7 \times 365}{60} \right)^{0.25} \times 1ns = 97.86ps \]

So the total Reduction is

\[ \frac{7 \text{ year guardband} - \left( 60 \text{ day guardband} + 20 \text{ps} \right)}{7 \text{ year guardband} - 250 - (97.86 + 20)} \times 100\% = 53\% \]

And we can see the value is got in the table. We add an additional 20ps to ensure reliable operation of our aging sensor design. Negative numbers in Table 5.4 correspond to the cases in which circuit failure prediction using our aging sensor design is not beneficial in reducing the guardband. In principle, a shorter initial guardband implies better performance (e.g., 1 day instead of 15 days). However, that requires very frequent monitoring and periodic on-line self-test to ensure that enough circuit paths have been exercised.

H. Self correction options

The next problem is about how to analyze these data detected by the sensor and how to do self-correction. There are mainly three methods:

1) Variable clock frequency:
   The clock frequency may be reduced over time depending on the actual usage and aging of a chip in the field;

2) Adaptive body bias:
   Since the primary effect of NBTI is to increase the magnitude of threshold voltage of a PMOS transistor, forward biasing the source-body junction can compensate for the change. Adaptive body bias has been shown to be effective for compensating die-to-die parameter variations and for reducing power in high-performance processors. Body bias has also been used for leakage improvement and process compensation. The tradeoff involved in using forward body bias for aging is the decrease in magnitude of threshold voltages for PMOS transistors that have not aged. This will result in increased leakage power;

3) Adaptive power supply:
   Another approach is to increase the voltage supply whenever the stability checker circuit detects a violation. It must be ensured that the increased supply is within the Vmax limit set by the technology. Increasing the supply voltage also unnecessarily speeds up parts of the circuit that have not aged and also accelerates aging.

VI Summary

This paper mainly talks about three aging mechanism effects: BTI, HCI and TDDB. And then we mainly talk about the CFP method for the NBTI-induced transistor aging problem.

Circuit failure prediction provides an effective way of overcoming reliability challenges in scaled-CMOS technologies. Effective failure prediction mainly consists of effective and efficient sensors design and self-correction. New aging sensor designs combining the concepts of pre-sampling and stability checking enables efficient and effective failure prediction at a very low cost. And from the simulation part we can see the system performance is actually improved because failure prediction assisted by the new aging sensor designs enable close to best-case design instead of traditional worst-case design.

REFERENCES