Optimized Circuit Failure Prediction for Aging: Practicality and Promise

Mridul Agarwal, AMD
Varsha Balakrishnan, Arizona State Univ.
Anshuman Bhuyan, Stanford Univ.
Kyunglok Kim, Stanford Univ.
Bipul C. Paul, Stanford Univ.
Wenping Wang, Arizona State Univ.
Bo Yang, Arizona State Univ.
Yu Cao, Arizona State Univ.
Subhasish Mitra, Stanford Univ.

ABSTRACT

Circuit failure prediction is used to predict occurrences of circuit failures, during system operation, before errors appear in system data and states. This technique is applicable for overcoming major scaled-CMOS reliability challenges posed by aging mechanisms such as Negative-Bias-Temperature-Instability (NBTI). This is possible because of the gradual nature of degradation associated with such aging mechanisms. Circuit failure prediction uses special on-chip circuits called aging sensors. In this paper, we experimentally demonstrate correct functionality and practicality of two flavors of flip-flop designs with built-in aging sensors using 90nm test chips. We also present an aging-aware timing analysis technique to strategically place such flip-flops with built-in aging sensors at selective locations inside a chip for effective circuit failure prediction. This aging-aware timing analysis approach also minimizes the chip-level area impact of such aging sensors. Results from two 90nm designs demonstrate the practicality and effectiveness of optimized circuit failure prediction with overall chip-level area impact of 2.5% and 0.6%.

1. INTRODUCTION

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. *Negative Bias Temperature Instability* or *NBTI* is a leading aging mechanism which limits circuit life time [Chen 03]. NBTI is prominent in PMOS transistors, and can shift (degrade) the PMOS threshold voltage by more than 50mV over ten years. This translates to more than 20% degradation in circuit speed [Shroder 03, Borkar 06]. Possible options to ensure reliable system performance in the presence of circuit aging are:

1. Improve process technology: While this option is attractive, process scaling is approaching physical and manufacturing limits.

2. Worst-case speed guardbands: This approach slows down the clock frequency of a chip (during design or test) based on the worst degradation that **may** occur due to the worst combination of temperature, voltage and workload over its useful lifetime. This approach is very pessimistic because most of the chips may not be stressed to worst levels in the field. Moreover, there is significant concern about increased circuit aging with CMOS scaling [Kuhn 08].

3. On-line circuit failure prediction and self-healing: *Circuit failure prediction* [Agarwal 07, Mitra 08a, 08b] predicts the occurrence of a circuit failure **before** errors actually appear in system data and states. This is in contrast to traditional error detection where a failure is detected **after** errors appear in system data and states. On-line circuit failure prediction is a promising solution to cope with circuit aging in an adaptive fashion for robust systems with built-in self-healing.

The basic principle behind circuit failure prediction is to collect information about the evolution of various system parameters over time, and to analyze the collected data to predict failures. The information is collected concurrently during normal system operation and / or during periodic on-line self-test. System parameters that we focus on in this paper involve timing characteristics of logic signals because they are effective in predicting circuit aging. This is because of the gradual delay degradation associated with circuit aging.

High-confidence circuit failure prediction for aging can be implemented in two ways (or a combination of both):

1. Concurrently during system operation using special circuits called *circuit failure prediction sensors* or *aging sensors* or *monitoring blocks*;

2. Thorough on-line self-tests assisted by aging sensors and / or clock control.

Traditional process monitors (e.g., ring oscillators) and temperature sensors are inadequate for circuit failure prediction targeting NBTI-induced transistor aging [Agarwal 07]. This is because the amount of aging strongly depends on the workload of a chip. The workload of process monitors can be very different from the actual workload experienced by various circuit blocks in a chip.

In [Agarwal 07], we introduced design of special flip-flops with integrated circuit failure prediction sensors for NBTI aging. As explained in [Agarwal 07], chip-level dynamic power impact of such flip-flops is very small since the associated circuit failure prediction sensors need to be activated only for a very small fraction of time (since NBTI aging is a slow process). However, the cell-level area of such a flip-flop can be 2.5 times the area of a conventional flip-flop. Since existing scan chains are reused to read out circuit failure prediction results, there is very little impact on global routing. Hence, flip-flops with built-in circuit failure prediction sensors must be inserted in carefully selected locations in order to maximize the effectiveness of circuit failure prediction while minimizing chip-level area impact.

The major contributions of this paper are:

1. We present experimental results to demonstrate the practicality of using flip-flops with integrated circuit failure prediction sensors (from [Agarwal 07]) using 90nm test chips.

2. Based on physical modeling of NBTI-induced aging, we introduce a new analysis method called *Maximum-Dynamic-Stress (MDS)* to estimate the maximum degradation of circuit timing under NBTI aging. This MDS-based simulation method accurately predicts a practical bound of circuit aging. Moreover, it helps identify critical regions of a circuit under NBTI, avoiding expensive computation and enables fast and safe reliability prediction.

3. We use the MDS analysis technique for insertion of flip-flops with integrated circuit failure prediction sensors at strategic locations inside a chip. This joint optimization approach maximizes the effectiveness of circuit failure prediction while minimizing area cost. We demonstrate the practicality and effectiveness of our overall approach using two 90nm designs – an Ethernet controller design and the OpenRISC processor design. The area impact of cir-

1-4244-4203-0/08/$20.00 ©2008 IEEE

cuit failure prediction sensors for these two designs, after optimized insertion, are 2.5% and 0.6%, respectively.

The outline of the rest of the paper is as follows: Section 2 presents the concept of circuit failure prediction (previously published in [Agarwal 07]), and demonstrates this concept using 90nm test chips; Section 3 describes the MDS based simulation method for the identification of critical FFs that require flip-flops with integrated circuit failure prediction sensors (described in Sec. 2). Section 4 demonstrates the reduction in chip-level area costs of circuit failure prediction sensors using the MDS based aging estimation method of Sec. 3. Section 5 concludes this paper.

2. CIRCUIT FAILURE PREDICTION

2.1 Concept

One application of circuit failure prediction in the context of transistor aging is to eliminate worst-case speed guardbands. Such an approach enables designs with close to best-case performance unlike traditional design techniques that impose worst-case speed guardbands [Agarwal 07]. Instead of introducing a worst-case aging guardband over the entire lifetime of a chip (e.g., 7 to 10 years for enterprise systems), the system starts off with a small timing guardband which guarantees correct circuit operation even with worst aging over a short period of time, e.g., 15 days (this period of time can vary depending on system design constraints). We refer to this timing guardband as the *guardband interval*. During this period, circuit failure prediction sensors collect data about relative aging of various circuit paths. At the end of 15 days, the data is analyzed to check whether there has been enough aging that requires adjustment of the guardband interval. If yes, the system adapts itself based on its operation history by adjusting the guardband based on the estimated amount of actual aging using special self-healing techniques (e.g., adjustment of supply voltage, threshold voltage, speed) [Agarwal 07].

A major component of circuit failure prediction for transistor aging is the design of an aging-resistant circuit failure prediction sensor (we will interchangeably use the terms circuit failure prediction sensor and aging sensor) for accurate on-line aging prediction. Figure 2.1 shows the working principle of such sensors in a digital system with rising edge-triggered flip-flops [Agarwal 07]. This technique modifies a standard flip-flop by inserting a “monitoring block” (Fig. 2.1a) which records “significant” shifts in the delay of the combinational logic whose output is connected to the data input of that flip-flop. The task of the monitoring block is to detect signal transitions at the combinational logic output during the guardband interval (denoted as T_g in Fig. 2.1b). Signal transitions during T_g imply that one or more paths in the combinational logic have aged enough to creep into the guardband interval. Note that, the flip-flop still continues to capture correct values unlike traditional error detection.

As explained in [Agarwal 07], there are two ways to design the monitoring block in Fig. 2.1a: 1. Stability checking based (Fig. 2.2); and, 2. Pre-sampling based (Fig. 2.3).

Figure 2.2b shows the design of a stability checker based aging sensor design (taken from [Agarwal 07]). 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. The delay element introduces a delay of ‘$T_{clk}/2 - T_g$’ (assuming 50% duty cycle of *Clock*). Here, Tclk is the clock period and T_g is the guardband interval (Fig. 2.1b). This guarantees that transistors T3 and T4 (and transistors T7 and T8) are both on during the guardband interval T_g. During the guardband interval, PMOS transistors T1 and T5 are turned off. Stability Checker output 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, i.e., the guardband is violated. The output latch in Fig. 2.2b is responsible for latching the stability checker output.

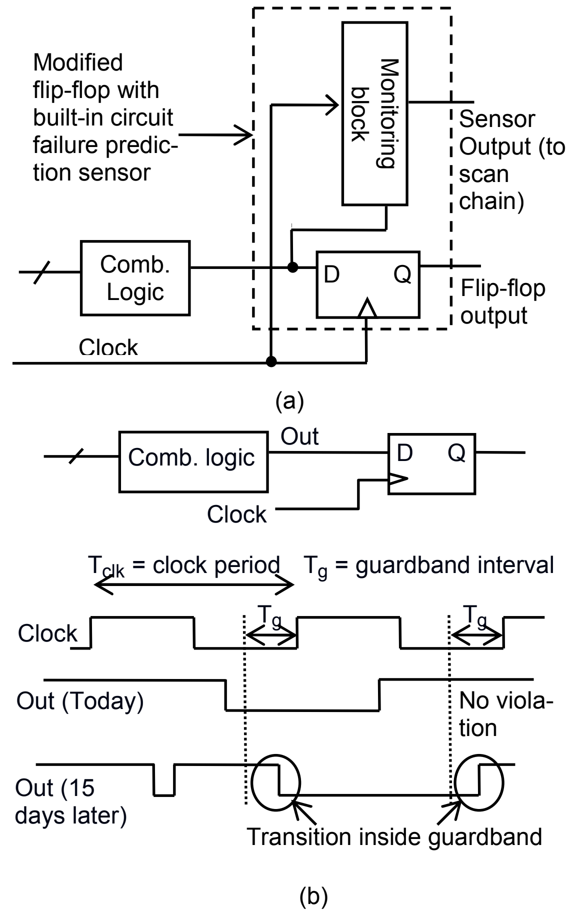

Figure 2.1. Circuit failure prediction for aging. (a) Flip-flop with built-in circuit failure prediction (aging) sensor. (b). Working principle.

An alternative aging sensor design based on pre-sampled outputs shown in Fig. 2.3. It pre-samples the output of the combinational logic right before the guardband interval and then compares it with the actual value sampled at the clock edge [Agarwal 07]. The delay of the delay element in Fig. 2.3 is the same as the guardband interval T_g. This sensor design can be prone to invalidation due to hazards. For example, a static hazard inside the guardband interval may not get reported by this sensor.

It is extremely important to ensure that the aging sensors themselves do not age significantly. For the designs in Fig. 2.2 and 2.3, the delay elements are especially prone to aging. Aging resistant delay elements, described in [Agarwal 07], overcome this problem (not repeated in this paper).

1-4244-4203-0/08/$20.00 ©2008 IEEE

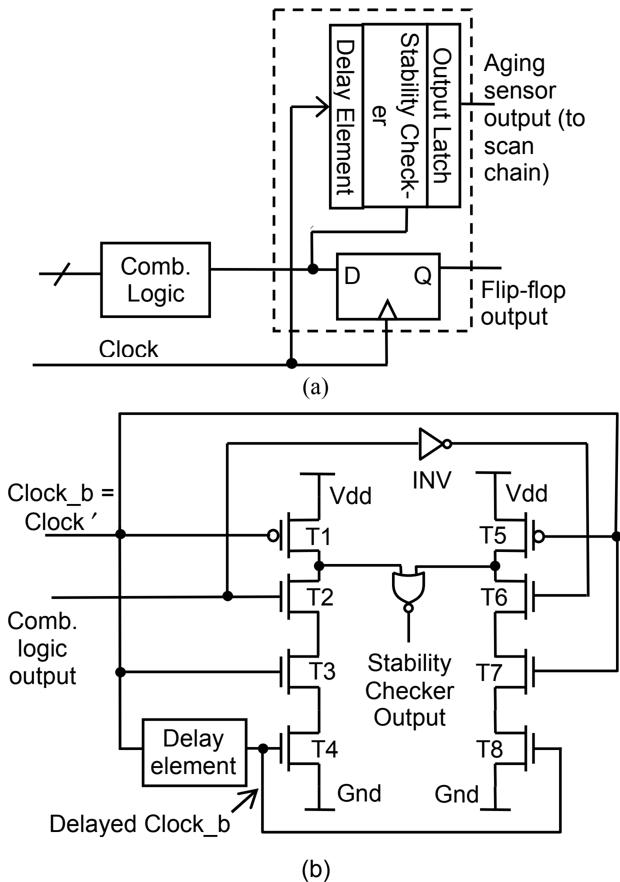

Figure 2.2. Stability checking based aging sensor. (a) Block diagram. (b) Transistor-level schematic.

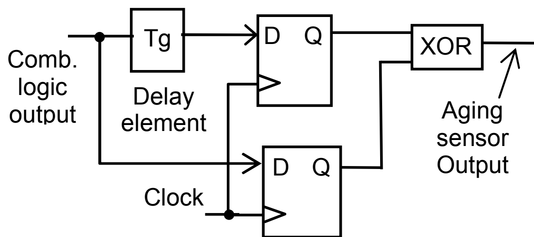

Figure 2.3. Pre-sampling based aging sensor design.

2.2 Test Chip Demonstration

We designed a 90nm test chip using a commercial CMOS process to demonstrate the functionality of the circuit failure prediction sensors on silicon. Figure 2.4 shows the layout of the test chip. The test chip consists of two main blocks, CFP Block 1 and CFP Block 2 (CFP stands for circuit failure prediction), to test Aging Sensor 1 (Fig. 2.2) and Aging Sensor 2 (Fig. 2.3), respectively.

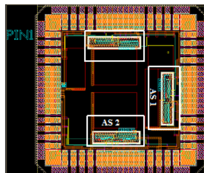

Figure 2.4. Test chip layout.

2.2.1 CFP Block 1 – Aging Sensor 1

Figure 2.5 shows a partial schematic of CFP Block 1. It consists of two combinational circuit blocks designed to run at the frequency range of 50-100 MHz. Each combinational block consists of seven inverter chains of different delays.

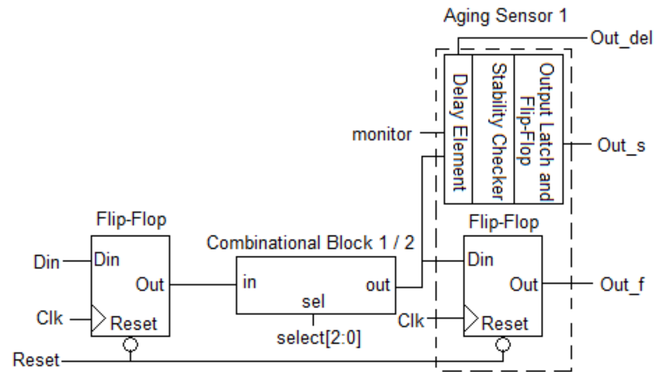

Figure 2.5. CFP Block 1 design.

Aging of combinational logic is primarily emulated using the inverter chains in combinational logic blocks. In addition, we provide experimental results obtained from stress testing. The inverter chains are designed to have different delays such that, for a given clock frequency, signal transitions at inverter chain outputs occur outside the guardband interval for some inverter chains, and inside the guardband interval for the others. (Examples of signal transitions inside and outside a guardband interval are shown in Fig. 2.1b). The precise width of the guardband interval is not important from the test chip standpoint (since our main focus is on the functionality of aging sensors). Hence, we chose the guardband intervals to be around 15-20% of the designed clock frequency of the circuit (around 50 MHz for Combinational Block 1 and 100 MHz for Combinational Block 2). In addition, we report experimental data on the resolution of aging sensors. The different inverter chains are selected through a multiplexer. The output of the aging sensor reports logic '1' when the output transition of an inverter chain occurs inside the guardband interval. The output latch of Aging Sensor 1 (Fig. 2.2a) is connected to a flip-flop with output 'Out_s.'

The input/output (I/O) pins of the testchip were carefully selected to maximize signal observability after fabrication. For each CFP block, there are three output signals (Out_del, Out_s and Out_f), two input signals (Clk and Din) and five control signals (select[2:0], monitor and reset). The 'monitor' signal is used to activate the operation of aging sensors when desired (Use of such 'monitor' signal to improve aging resistance of aging sensors is discussed in [Agarwal 07]). The reset signal is used to initialize all flip-flops to logic '0'. The different inverter chains are selected using the three signals in 'select[2:0]'. 'Out_f' is the output of the flip-flop (clocked by the Clk signal) connected to the combinational block. The output of the delay element, used for the aging sensor (Fig. 2.2, details of the delay element design appear in [Agarwal 07]), is also strobed by a flip-flop with output 'Out_del' (clocked by the Clk signal).

2.2.2 CFP Block 2 – Aging Sensor 2

CFP Block 2 (Fig. 2.6) is similar to CFP Block 1. Aging Sensor 2 (Fig. 2.3) is used instead of Aging Sensor 1.

1-4244-4203-0/08/$20.00 ©2008 IEEE

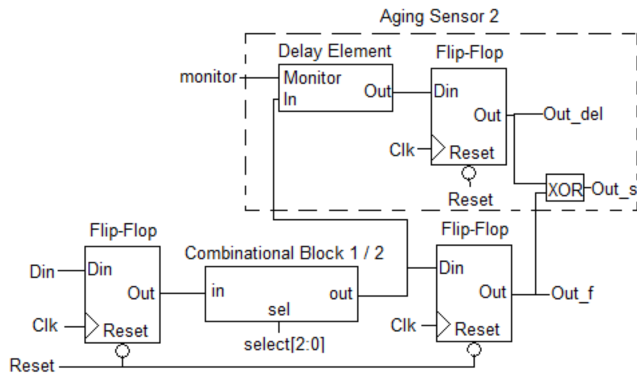

Figure 2.6. CFP Block 2 design.

2.2.3 Measurement Techniques and Results

We estimated the delay of each inverter chain by applying proper multiplexer select inputs, varying the clock period, and observing the flip-flop output 'Out_f'. The minimum clock period at which the correct result (transition at the output of an inverter chain) is obtained is considered the delay of the inverter chain. Note that, this measured delay also includes the setup time of the destination flip-flop. Tables 2.1 and 2.2 list delays of various inverter chains in CFP Block 1 (associated with Aging Sensor 1) and CFP Block 2 (associated with Aging Sensor 2), respectively, of one arbitrarily selected chip. Each inverter chain delay was measured several times to check for consistency in measured results. These measurements were done at short intervals of time, e.g., within few minutes, and we did not see any significant variations in measured results. The delays of the delay elements were also measured in a similar way by observing the flip-flop output 'Out_del'. The measured delays of the delay elements of Combinational Block 1 for CFP Block 1 and CFP Block 2 are 5.04 ns and 3.88 ns, respectively.

When an output transition of an inverter chain occurs within the guardband interval, the output of the corresponding aging sensor reports logic '1'. We observe the aging sensor outputs (Out_s) for all seven inverter chains in each combinational circuit block of each CFP block. Due to different inverter chain delays, only some of them have output transitions inside the guardband interval. Hence, the corresponding aging sensor reports a '1' only when those inverter chains are selected.

Table 2.1. Inverter Chain Delays in CFP Block 1 (for Aging Sensor 1).

Inverter Chain Identifier	Delay (ns)	
	Combinational Block 1	Combinational Block 2
1	12.960	6.020
2	13.828	6.536
3	14.240	6.780
4	15.024	7.004
5	15.064	7.020
6	15.352	7.064
7	16.340	7.108

Table 2.2. Inverter Chain Delays in CFP Block 2 (for Aging Sensor 2).

Inverter Chain Identifier	Delay (ns)	
	Combinational Block 1	Combinational Block 2
1	13.274	5.038
2	14.172	5.488
3	14.356	5.612
4	14.472	5.926
5	14.648	6.040
6	15.172	6.088
7	15.382	6.184

Due to the nature of the design of Aging Sensor 1, the width of the guardband interval changes with varying clock period. However, that is not the case for Aging Sensor 2. Table 2.3 shows the expressions for the positions of the rising edge of the guardband interval for constant delay (T_d) of the delay element – the reference point is the rising edge of the clock signal which appears right before the rising edge of the guardband interval. (Earlier in Sec. 2, we already discussed that for constant guardband interval T_g, the delays of the delay elements for Aging Sensors 1 and 2 are $T_{clk}/2 - T_g$ and T_g, respectively. Table 2.3 is another way to represent the same information but for constant delay element instead of constant guardband interval). For example, consider an inverter chain of delay (T_{inv}) 14 ns, delay element delay (T_d) of 5 ns, and a clock period (T_{clk}) of 20 ns. Using the expressions in Table 2.3, the position of the rising edge of the guardband interval (with respect to the rising edge of the clock signal before the guardband interval) is at 15 ns for both sensors, and the inverter chain transition occurs outside the guardband interval (Figs. 2.7 and 2.9). The width of the guardband interval is 5 ns for both aging sensors. However, if the clock period (T_{clk}) is reduced to 15 ns, the absolute position of the rising edge of the guardband interval (with respect to the rising edge of the clock signal before the guardband interval) will be at 12.5 ns for Aging Sensor 1, and 10 ns for Aging sensor 2 (Figs. 2.8 and 2.10). We notice the width of the guardband interval associated with Aging Sensor 1 is also reduced from 5 ns to 2.5 ns. The inverter chain transition will now occur inside the guardband interval for both aging sensors. Thus, by varying the clock frequency, we see that we can control the position of the transition with respect to the guardband interval rising edge.

For each inverter chain, we swept the clock frequency (in step size of 100 ps) and recorded the output of the corresponding aging sensor to generate Shmoo tables (Tables 2.4 - 2.7). An entry with 'Y' indicates that the aging sensor recorded a transition inside the guardband interval ('X' indicates no detection is recorded). The staircase-like behaviors of the Shmoo tables indicate correct operations of the aging sensors.

We repeated the above measurements at different supply voltages. For each supply voltage, we measured the clock period at which the output transition of an inverter chain occurred just inside the guardband interval (referred to as the *trip point* of that inverter chain). Since the delay of an inverter chain increases with reduced supply voltage, the corresponding clock period must also increase

1-4244-4203-0/08/$20.00 ©2008 IEEE

to find the trip point for that inverter chain (follows from Table 2.3). Figures 2.11 and 2.12 show the trip points associated with various inverter chains over varying voltages and demonstrate correct behaviors of aging sensors.

Table 2.3. Absolute Position of Guardband Interval Rising Edge with respect to the Rising Edge of the Clock Signal before the Guardband Interval.

Aging Sensor 1	*Aging Sensor 2*
$T_{clk}/2 + T_d$	$T_{clk} - T_d$

T_{clk} – Clock period, T_d – Delay (Delay element)

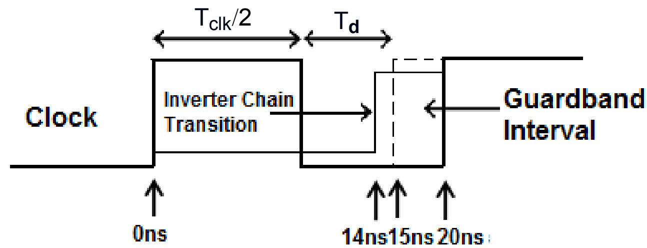

Figure 2.7. Position of the rising edge of the guardband interval with respect to the rising edge of the clock signal before the guardband interval for Aging Sensor 1. (Illustrated example assumes at 20 ns clock period).

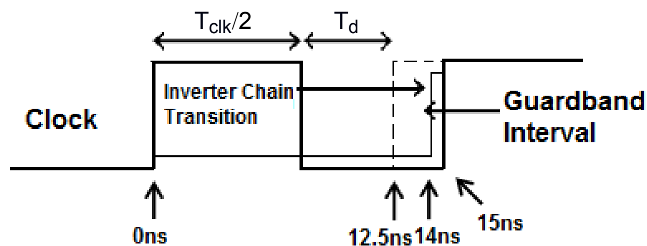

Figure 2.8. Position of the rising edge of the guardband interval with respect to the rising edge of the clock signal before the guardband interval for Aging Sensor 1. (Illustrated example assumes 15 ns clock period).

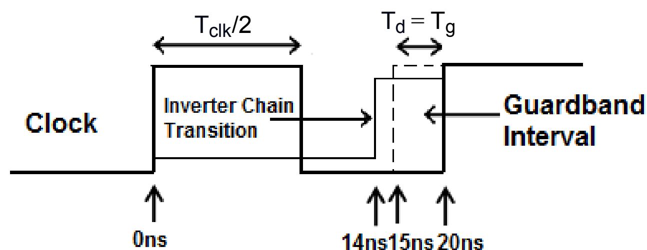

Figure 2.9. Position of the rising edge of the guardband interval with respect to the rising edge of the clock signal before the guardband interval for Aging Sensor 2. (Illustrated example assumes 20 ns clock period).

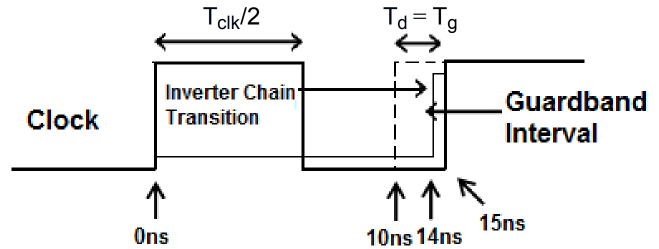

Figure 2.10. Position of the rising edge of the guardband interval with respect to the rising edge of the clock signal before the guardband interval for Aging Sensor 2. (Illustrated example assumes 15 ns clock period).

Table 2.4. Shmoo Plot - Combinational Block 1 of CFP Block 1 (Corresponding to Aging Sensor 1).

INVERTER CHAINS at 1 V Supply

CLOCK PERIOD (ns)	1	2	3	4	5	6	7
18.1	X	Y	Y	Y	Y	Y	Y
18.2	X	Y	Y	Y	Y	Y	Y
18.3	X	X	Y	Y	Y	Y	Y
19.0	X	X	Y	Y	Y	Y	Y
19.1	X	X	X	Y	Y	Y	Y
19.2	X	X	X	Y	Y	Y	Y
19.3	X	X	X	X	Y	Y	Y
19.8	X	X	X	X	Y	Y	Y
19.9	X	X	X	X	X	Y	Y
21.0	X	X	X	X	X	Y	Y
21.1	X	X	X	X	X	X	Y
21.5	X	X	X	X	X	X	Y
21.6	X	X	X	X	X	X	X

Table 2.5. Shmoo Plot - Combinational Block 2 of CFP Block 1 (Corresponding to Aging Sensor 1).

INVERTER CHAINS at 1 V Supply

CLOCK PERIOD (ns)	1	2	3	4	5	6	7
9.0	X	Y	Y	Y	Y	Y	Y
9.1	X	Y	Y	Y	Y	Y	Y
9.2	X	X	Y	Y	Y	Y	Y
11.4	X	X	Y	Y	Y	Y	Y
11.5	X	X	X	Y	Y	Y	Y
11.6	X	X	X	X	Y	Y	Y
11.9	X	X	X	X	Y	Y	Y
12.0	X	X	X	X	X	Y	Y
12.1	X	X	X	X	X	Y	Y
12.2	X	X	X	X	X	X	Y
12.3	X	X	X	X	X	X	Y
12.6	X	X	X	X	X	X	Y
12.7	X	X	X	X	X	X	X

1-4244-4203-0/08/$20.00 ©2008 IEEE

Table 2.6. Shmoo Plot - Combinational Block 1 of CFP Block 2 (Corresponding to Aging Sensor 2).

INVERTER CHAINS at 1 V Supply

CLOCK PERIOD

(ns)	1	2	3	4	5	6	7
17.0	Y	Y	Y	Y	Y	Y	Y
17.1	X	Y	Y	Y	Y	Y	Y
17.9	X	Y	Y	Y	Y	Y	Y
18.0	X	X	Y	Y	Y	Y	Y
18.2	X	X	Y	Y	Y	Y	Y
18.3	X	X	X	X	Y	Y	Y
18.5	X	X	X	X	Y	Y	Y
18.6	X	X	X	X	X	Y	Y
19.1	X	X	X	X	X	Y	Y
19.2	X	X	X	X	X	X	Y
19.5	X	X	X	X	X	X	Y
19.6	X	X	X	X	X	X	X
19.7	X	X	X	X	X	X	X

Table 2.7. Shmoo Plot - Combinational Block 2 of CFP Block 2 (Corresponding to Aging Sensor 2).

INVERTER CHAINS at 1 V Supply

CLOCK PERIOD

(ns)	1	2	3	4	5	6	7
6.3	Y	Y	Y	Y	Y	Y	Y
6.4	X	Y	Y	Y	Y	Y	Y
6.8	X	Y	Y	Y	Y	Y	Y
6.9	X	X	Y	Y	Y	Y	Y
7.0	X	X	X	Y	Y	Y	Y
7.1	X	X	X	Y	Y	Y	Y
7.2	X	X	X	Y	Y	Y	Y
7.3	X	X	X	Y	Y	Y	Y
7.4	X	X	X	X	Y	Y	Y
7.5	X	X	X	X	X	Y	Y
7.6	X	X	X	X	X	X	Y
7.7	X	X	X	X	X	X	X

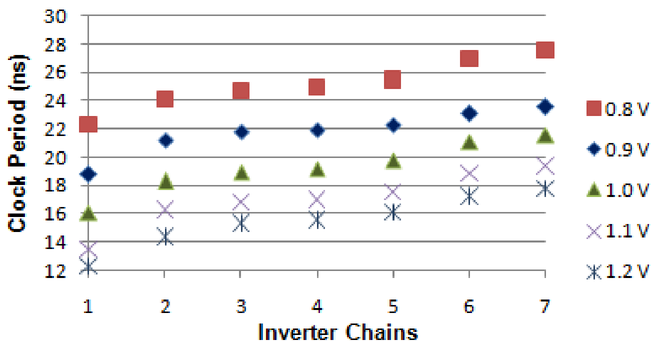

Figure 2.11. Trip points - Combinational Block 1 of CFP Block 1 (Corresponding to Aging Sensor 1).

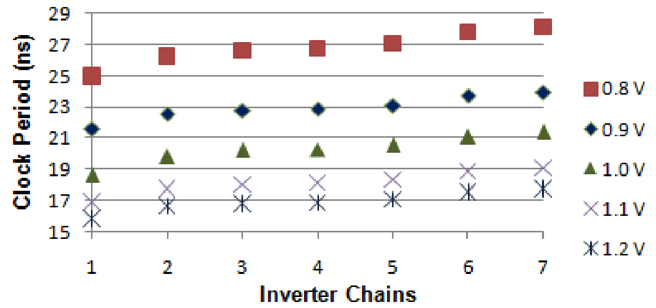

Figure 2.12. Trip points - Combinational Block 1 of CFP Block 2 (Corresponding to Aging Sensor 2).

Next, we calculate the resolution of Aging Sensor 1. The *resolution* is the minimum overlap required between the signal transition at the inverter chain output and the rising edge of the guardband interval for the aging sensor to start reporting signal transition (Fig. 2.13). To calculate the resolution of the aging sensor, three measurements are required – delay of the inverter chain (T_{inv}), delay of the delay element (T_d) and the clock period at which the aging sensor first reports a transition (T_{detect}). The resolution can then be calculated using the following formula: Resolution = $T_{inv} - (T_d + T_{detect}/2)$. Note that, both T_{inv} and T_d have setup times incorporated in the measurement but they cancel out to first order. One drawback of this approach is that it is prone to errors due to measurement uncertainties. One must also ensure that, while measuring the resolution, the output transition of the inverter chain must be a rising transition. This ensures that we measure the worst-case resolution. Figure 2.13 provides a graphical illustration of resolution measurement. It assumes that the clock duty cycle is 50%.

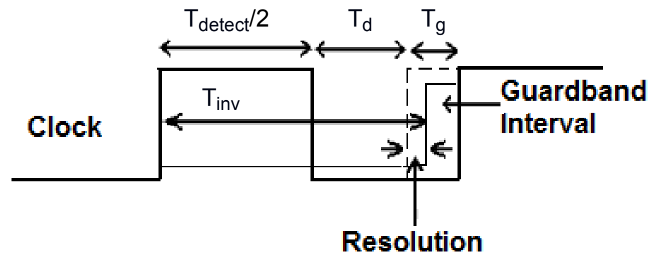

Figure 2.13. Resolution Measurement Principle for Aging Sensor 1.

Table 2.8 lists the resolution of Aging Sensor 1 for various inverter chains in Combinational Block 1 of CFP Block 1 of an arbitrarily selected chip. Each measurement was repeated several times over a short interval of time to check for consistency of results. Note that, the negative values of the resolution observed in some cases were due to the measurement uncertainties. Especially, since the clock signal was generated externally (and not on-chip with proper skew control) and the resolution was obtained indirectly by measuring T_{inv}, T_d and T_{detect}, the uncertainties in measurement equipment played a crucial role. Nevertheless, the small resolution value of the aging sensor demonstrates its effectiveness in successfully detecting signal transitions inside the guardband interval. We performed the above resolution experiment across multiple chips (10) for two arbitrarily selected inverter chains. Each measurement was performed multiple times for each chip to check for consistency of results. Table 2.9 presents the results.

1-4244-4203-0/08/$20.00 ©2008 IEEE

Table 2.8. Resolution results: Combinational Block 1 of CFP Block 1.

Inverter Chain Identifier	Logic Delay (ns)	Delay Element (ns)	T_{detect} (ns)	Resolution (ps)
1	12.960	5.04	15.768	36
2	13.828	5.04	17.624	-24
3	14.240	5.04	18.344	28
4	15.024	5.04	19.808	80
5	15.064	5.04	20.152	-52
6	15.352	5.04	20.816	-96
7	16.340	5.04	22.400	100

Table 2.9. Resolution Results for Aging Sensor 1 over Several Chips.

Chip id	Inverter Chain 1	Inverter Chain 2
1	-30 ps	1 ps
2	90 ps	29 ps
3	0 ps	-21 ps
4	-36 ps	0 ps
5	65 ps	-77 ps
6	110 ps	-16 ps
7	65 ps	-43 ps
8	-6 ps	24 ps
9	-18 ps	121 ps
10	24 ps	-3 ps
Mean	26.4 ps	1.5 ps
Standard Deviation	52.5 ps	62.8 ps

We performed stress experiments by stressing some arbitrarily selected chips at 1.5V. The monitor signal was turned off, making the delay element resistant to aging (details in [Agarwal 07]). We performed delay measurements at the nominal voltage (1 V) and then stressed the chips at 1.5 V, and repeated this procedure several times over a period of 20 days. Table 2.10 lists the delays of an arbitrarily selected inverter chain and the corresponding delay element over this period of 20 days (for one of the stressed chips). Three points to keep in mind about Table 2.10: 1. Measurement uncertainties due to the measurement setup can lead to variations of 50-100ps; 2. Each measurement was taken at an interval of 3 days (except for the 4th measurement); and, 3. Between the 3rd and 4th entries, the chip was allowed to rest (the chip was removed from the socket) for about 6 days (between day 7 and 13) before it was stressed again. Table 2.10 indicates that there is a clear delay shift of the inverter chain. However, there is no significant shift in the delay of the delay element (since the Monitor signal was held at logic 0). We also verified that the increased delay of the inverter chain in Table 2.10 (whose transition originally occurred outside the guardband interval) resulted in signal transitions inside the guardband interval. As a result, the aging sensor output, which originally reported logic '0' for this inverter chain (for a given clock period), reported a logic '1' after stress (for the same clock period).

Table 2.10. Delay measurements over a period of 20 days. The chip was allowed to rest between Day 7 and 13 (by removing it from the test socket).

Day	Inverter Chain Delay (ns)	Delay Element Delay (ns)
1	13.988	5.412
4	14.236	5.516
7	14.244	5.492
13	14.164	5.384
16	14.556	5.400
19	14.600	5.388

3. CIRCUIT AGING ANALYSIS

Aging sensors, described and demonstrated in Sec. 2, can affect chip area unless they are inserted carefully at strategic locations inside a chip. To minimize this area impact, judicious use of aging sensors is essential during design phase. For circuit failure prediction using aging sensors, a naïve approach is to insert aging sensors at all flip-flops with at least one circuit path (ending at that flip-flop) whose timing slack is less than or equal to a certain guardband interval (as discussed in Sec. 2.1). This approach is highly pessimistic because not all circuit paths in a design will degrade by the same amount [Wang 07a]. This is because NBTI-induced degradation is a strong function of the percentage of time a PMOS transistor is on (often referred to as *duty cycle* in NBTI literature which should not be confused with clock duty cycle). In this section, we present a *Maximum-Dynamic-Stress (MDS)* simulation technique based on a compact NBTI model to determine the upper bound on the degradation of each gate in a design. By combining MDS and static timing analysis, we determine the set of flip-flops that require aging sensors for circuit failure prediction. As a result, the chip-level area overhead of aging sensors is minimized enabling cost-effective circuit failure prediction.

3.1 Degradation Model under NBTI Effect

For a PMOS transistor, there are two phases of NBTI depending on its bias condition – stress phase and recovery phase. When V_g=0 (i.e., V_{gs}= -V_{DD}), positive interface charges accumulate over time with H-species diffusing towards the gate. This phase is usually referred as *stress* phase, and the time during which the device is under V_g=0 bias condition is defined as *stress time*. When V_g=V_{DD} (i.e., V_{gs}=0), holes are not present in the channel and hence, no new interface charges are generated. Instead, H-species diffuse back and anneal the interface charges. As a result, the number of interface charges is reduced during this stage, and some part of the NBTI degradation is recovered. We usually refer to this phase as *recovery* phase. If a PMOS device is always under stress, it is described as *static NBTI*. Otherwise, when both stress and recovery exist during active circuit operation, we refer to it as *dynamic NBTI*, in which case the amount of degradation depends on the percentage of time a PMOS is under stress (i.e., duty cycle). Figure 3.1 shows the threshold voltage degradation under dynamic NBTI, with analytical model prediction verified by 90nm silicon data [Wang 07b]. As

1-4244-4203-0/08/$20.00 ©2008 IEEE

shown in Fig. 3.1, there is a sudden change in threshold voltage (ΔV_{th}) in the beginning of the recovery phase. This sudden drop in ΔV_{th} can be attributed to fast H-diffusion in the gate dielectric or trapping/detrapping [Huard 06]. After the sudden drop, the recovery of V_{th} happens slowly, which is limited by the slow H-diffusion in the poly-silicon gate [Wang 07b].

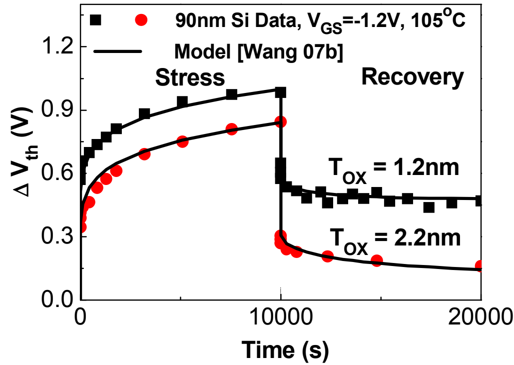

Figure 3.1. Threshold voltage degradation under dynamic NBTI.

Based on the short-term dynamic NBTI model in [Wang 07b], a long-term threshold voltage degradation model is derived as [Bhardwaj 06]:

$$\Delta V_{th} = \left[\sqrt{K_v^2 \cdot T_{clk} \cdot \alpha} \Big/ \left(1 - 2\sqrt[n]{\beta_t}\right) \right]^{2n} \quad (1)$$

where β_t and K_v are functions of voltage, temperature, total dynamic stress time, and switching activity [Bhardwaj 06]; T_{clk} is the clock cycle; α is duty cycle, which is defined as the probability in one clock cycle that a PMOS is in stress mode [Bhardwaj 06], and n is typically around 0.16 as a characteristic of NBTI effect. Since the recovery phase significantly affects the degradation, ΔV_{th} exhibits a clear dependence on α, as shown in Equation (1) and Fig. 3.2. In fact, during dynamic NBTI, the amount of V_{th} degradation is less than half of that in static NBTI (Fig. 3.2). Therefore, a correct dynamic NBTI model is necessary to accurately predict NBTI-induced degradation.

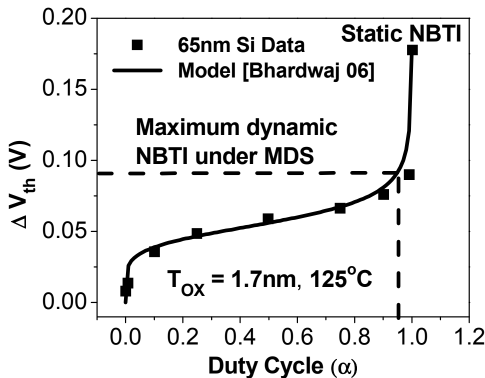

Figure 3.2. Dependence of V_{th} degradation on duty cycle.

Furthermore, Fig. 3.2 illustrates a sudden change in ΔV_{th} when duty cycle is close to 1, i.e., shifting from dynamic NBTI to static NBTI. When α increases from 0.5 to 0.9, ΔV_{th} only changes by 10%; but ΔV_{th} is abruptly doubled when α goes from 0.9 to 1 (static NBTI). This behavior has been observed in both silicon measurements and model predictions (Fig. 3.2) [Grasser 07, Huard 07]. The rapid ΔV_{th} change when α approaches 1 corresponds to the fast drop of ΔV_{th} in the beginning of the recovery phase (Fig. 3.1). Since a major portion of V_{th} degradation is recovered during this fast process, even a short period in the recovery phase is effective in recovering most of the degradation leading to a dramatic difference between dynamic and static NBTI. On the other hand, the dependence of dynamic NBTI on duty cycle is accounted for by the slow H-diffusion in the poly-silicon gate, i.e., the shallow portion of the recovery phase after the fast drop in Fig. 3.1.

Given a process technology, the long-term degradation model (Equation (1)) is further simplified as a power law of the total time in dynamic NBTI (i.e., t) and α: $\Delta V_{th} = b\, \alpha^n\, t^n$, where b is technology dependent and n is typically around 0.16 as a characteristic of NBTI effect. Since the degradation of gate delay caused by NBTI (Δt_{pi}) is linearly proportional to ΔV_{th} , we express it as:

$$\Delta t_{pi} = t_{p0i} \cdot \alpha^n \cdot t^n \quad (2)$$

where t_{p0i} is a parameter related to nominal gate delay that is defined between one input node of the gate and the gate output and n is the same as that in Equation (1). In the case of a gate with multiple input nodes (i.e., a multi-input gate), gate delay is calculated between each input node and the output.

3.2 Derivation of MDS

As discussed in Sec. 3.1, delay degradation caused by dynamic NBTI has a clear dependence on temperature, voltage, and duty cycle. We usually use high temperature and high voltage for an upper bound on degradation. However, such an approach will be overly pessimistic if we ignore the recovery effect and use α=1 (static NBTI) as a bound for duty cycle (Fig. 3.2). In this section, we present a *Maximum-Dynamic-Stress (MDS)* simulation technique that gives a simple and realistic estimation of the upper limit of gate delay degradation under dynamic NBTI.

The degradation rate of different circuit paths is pronouncedly different due to different switching activities and circuit topologies [Wang 07a]. However, the search for the exact maximum degradation is computationally expensive because of the extremely large space of duty cycles for each gate input node (input duty cycle). In the MDS method, we use duty cycle of 0.95 for each input node of a gate to calculate an upper bound of dynamic NBTI degradation. Since dynamic NBTI is a monotonic function of α, such a high value of α is close to the worst case of dynamic degradation. On the other hand, it avoids the pessimistic calculation of ΔV_{th} if static NBTI (α=1) is used. Therefore, the predicted gate delay is a practical bound of dynamic NBTI; it is also a tight bound, since ΔV_{th} only changes slightly when α increases from 0.5 to 0.95 due to the slow diffusion process in the recovery (Fig. 3.1).

Combined with Equation (2), the MDS method is used to predict the upper bound of delay degradation of each gate, where duty cycles of input node are set to 0.95. Figure 3.3 shows the delay degradation of various logic gates under different duty cycles. As expected, the dynamic degradation of gate delay reaches the maximum when α approaches 1. On the other hand, if α=1 (static NBTI) is used to calculate the delay degradation, we would overestimate

1-4244-4203-0/08/$20.00 ©2008 IEEE

the degradation by more than 100% due to the rapid change of ΔV_{th} as shown in Fig. 3.2. We therefore use this maximum-dynamic-stress (α=0.95) method to predict the maximum degradation rate of each gate in dynamic NBTI. The static degradation model (α=1) is only applied if the gate is not switching, e.g., at the standby mode.

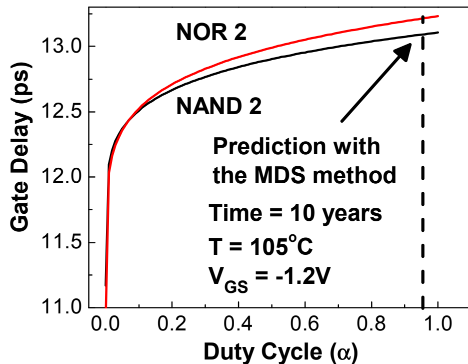

Figure 3.3. Dynamic NBTI-induced gate delay degradation under various duty cycles at the input node.

The gate-based MDS method is applied to each gate in a technology library to calculate its aging. The maximum delay degradation of each gate is generated by using Equation (2) and setting α of input node as the MDS value (α=0.95). Based on this library of "aged gates", we then use static timing analysis to predict the timing degradation of a given design.

The MDS based aging analysis method has a complexity comparable to gate-based static timing analysis, ignoring the complicated logic correlation among gates. The path degradation predicted by the MDS method is close to the upper bound since the computation is based on the maximum degradation of each gate. To demonstrate this, we perform a set of aging simulations using ISCAS benchmark circuits, with α's at the primary input nodes randomly distributed between 0 and 1. Depending on the topology of this circuit, the randomly assigned input duty cycles are propagated to each gate input to calculate the exact amount of dynamic degradation using equation (2). Figure 3.4 shows the distribution of the delay at critical paths after ten years for the ISCAS benchmark circuit Comp. ΔT_{max} is the maximum of circuit aging from these random simulations. As compared to ΔT_{max}, the bound predicted by the MDS method (ΔT_{ds}) is only 2.3% larger. Similar results have been obtained from simulations using other ISCAS benchmark circuits. These results demonstrate that the MDS method provides a safe and tight estimation of the maximum degradation.

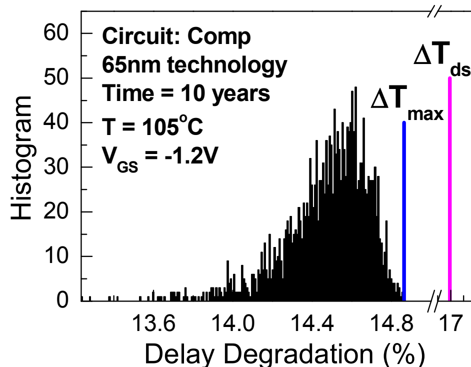

Figure 3.4. Histogram of delay degradation for various α's.

4. INTEGRATION OF CIRCUIT FAILURE PREDICTION AND MDS

4.1 Overview

Sections 2 and 3 demonstrated the practicality of two important concepts – Circuit Failure Prediction, and MDS-based simulation method to analyze circuit aging. This section combines these two concepts and demonstrates the cost-effectiveness of optimized circuit failure prediction by minimizing the number of aging sensors inserted in a design. We use aging-aware static timing analysis (STA) for this purpose.

Figure 4.1 illustrates the flow of our analysis. We synthesized our designs using standard Synopsys tools. Once we obtained the netlist, we used the MDS method to calculate the maximum degradation of each gate under dynamic NBTI. Next, we ran commercial tools (Synopsys Design Compiler and PrimeTime) and performed STA to identify the endpoint flip-flops that may violate setup time constraint under aging. These flip-flops need to be replaced with flip-flops with built-in aging sensors.

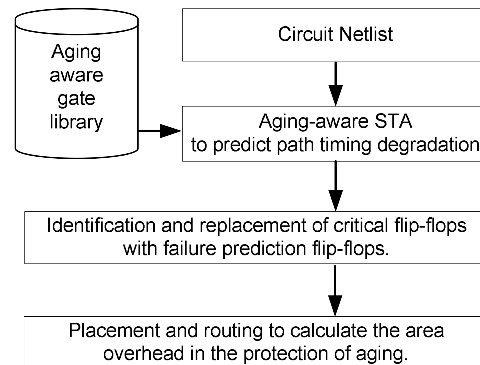

Figure 4.1. Flow of chip-level area overhead analysis for flip-flops with built-in aging sensors.

4.2 Design Examples

In order to demonstrate the effectiveness of our approach, we applied the above analysis approach to two designs from www.opencores.org: an Ethernet Controller and the OpenRISC processor. We used 90nm CMOS technology for both designs. The Ethernet Controller implements the MAC (Media Access Controller) portion of the Ethernet standard, connecting a 10/100Mbps Ethernet transceiver to a processor or other host. It is designed to run according to the IEEE 802.3 and 802.3u specifications. There are 2,352 flip-flops in the design. The OpenRISC processor is a 32-bit, 5-stage pipelined, in-order RISC processor with separate instruction/data caches and TLBs. It uses 1,366 flip-flops.

Tables 4.1 and 4.2 show the number of flip-flops that require built-in aging sensors before (Design I) and after (Design II) incorporating the aging analysis for the above two designs. For example, consider the Ethernet Controller design. The total number of flip-flops in the design is 2,352. Without aging-aware timing analysis, we chose 20% of clock period as the empirical margin for 10 year life time [Borkar 06]. Based on this margin, we insert aging sensors in flip-flops that have at least one circuit path connected to them that violate this timing slack (Design I). Without aging-aware STA analysis, 44% of the flip-flops need to be replaced with flip-flops with built-in aging sensors. That number significantly reduces to

1-4244-4203-0/08/$20.00 ©2008 IEEE

16% after performing the aging-aware STA analysis (Design II). Place and route results of the Ethernet Controller design at 90nm using Synopsys Astro (Table 4.1) show that area impact of aging sensors is significantly reduced to only 2.5% after performing aging-aware STA.

It is clear that the aging-aware STA enables optimized circuit failure prediction resulting in significant reduction in chip-level area costs associated with flip-flops with built-in aging sensors. As shown in Tables 4.1 and 4.2, the chip-level area impact of circuit failure prediction can be small, making it practical for overcoming the challenges associated with transistor aging in very large-scale designs. As discussed before and also in [Agarwal 07], the dynamic power impact of such aging sensors is very small because circuit failure prediction is activated very infrequently.

Table 4.1: Area Impact of Aging Sensors for the Ethernet Controller.

	No: of Flip-Flops with Aging Sensors	Area Overhead
Design I	1046 (44%)	7.0 %
Design II	379 (16%)	2.5 %

Table 4.2: Area Impact of Aging Sensors for OpenRISC.

	No: of Flip-Flops with Aging Sensors	Area Overhead
Design I	837 (61%)	0.7 %
Design II	713 (52%)	0.6 %

5. CONCLUSIONS

This paper establishes the concept of optimized circuit failure prediction for NBTI-induced transistor aging guided by a new aging-aware timing analysis technique. This optimization makes circuit failure prediction promising for future designs. The practicality of aging sensors for circuit failure prediction is demonstrated using 90nm test chips. The area costs of circuit failure prediction are significantly reduced, e.g., 2.5% and 0.6%, as demonstrated using actual 90nm designs. Future research directions include: 1. Evaluation of the necessity of on-line self-test techniques such as CASP and VAST [Inoue 08, Li 08] for accurate circuit failure prediction; 2. Use of on-line self-tests alone for low-cost circuit failure prediction (with minimum reliance on aging sensors); and, 3. Techniques for initiating appropriate self-healing actions upon circuit failure prediction.

ACKNOWLEDGEMENT

The authors acknowledge the support of the Center for Circuit and System Solutions (C2S2) and the Gigascale Systems Research Center (GSRC), two of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation program. This work is partially supported by NSF. The authors also thank UMC for support. The authors acknowledge the following people from Stanford University: Jason Hu, Prof. Boris Murmann, and the members of Prof. Mark Horowitz's research group, especially Valentin Abramzon, Zain Asgar and Dr. Don Stark.

REFERENCES

[Agarwal 07] M. Agarwal, et al., "Circuit Failure Prediction and Its Application to Transistor Aging," *IEEE VLSI Test Symp.*, 2007.

[Bhardwaj 06] S. Bhardwaj, et al., "Predictive Modeling of the NBTI Effect for Reliable Design," *IEEE Custom Integrated Circuits Conf.*, 2006.

[Borkar 06] S. Borkar, "Electronics beyond nano-scale CMOS," *ACM/IEEE Design Automation Conf.*, 2006.

[Chen 03] G. Chen, et al., "Dynamic NBTI of PMOS transistors and its impact on device lifetime," *International Reliability Physics Symp.*, 2003.

[Grasser 07] T. Grasser, et al., "Simultaneous extraction of recoverable and permanent components contributing to bias-temperature instability," *International Electron Device Meeting*, 2007.

[Huard 06] V. Huard, et al., "NBTI degradation: From physical mechanisms to modeling," *Microelectronics Reliability*, Jan 2006, vol. 46, no. 1, pages1-23.

[Huard 07] V. Huard, et al., "New characterization and modeling approach for NBTI degradation from transistor to product level," *International Electron Device Meeting*, 2007.

[Inoue 08] H. Inoue, Y. Li and S. Mitra, "VAST: Virtualization Assisted Concurrent Autonomous Self-Test," *Intl. Test Conf.*, 2008.

[Kuhn 08] K. Kuhn et al., "Managing Process Variation in Intel's 45nm CMOS Technology," *Intel Technology Journal*, June 2008.

[Li 08] Y. Li, S. Makar and S. Mitra, "CASP: Concurrent Autonomous Chip Self-Test using Stored Test Patterns," *Design Automation and Test in Europe*, 2008.

[Mitra 08a] S. Mitra, "Globally Optimized Robust Systems to Overcome Scaled CMOS Challenges," *Design Automation and Test in Europe*, 2008.

[Mitra 08b] S. Mitra, "Circuit Failure Prediction for Robust System Design in Scaled CMOS," *International Reliability Physics Symp.*, 2008.

[Shroder 03] D. K. Schroder, et al, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *Journal of Applied Physics*, 2003.

[Wang 07a] W. Wang, et al., "The impact of NBTI on the performance of combinational and sequential circuits," *Design Automation Conf.*, 2007.

[Wang 07b] W. Wang, et al., "Compact Modeling and Simulation of Circuit Reliability for 65nm CMOS Technology," *IEEE Trans. Device and Materials Reliability*, 2007.

1-4244-4203-0/08/$20.00 ©2008 IEEE