Efficient Observation Point Selection for Aging Monitoring

Liu, Chang; Kochte, Michael A.; Wunderlich, Hans-Joachim

Proceedings of the 21st IEEE International On-Line Testing Symposium (IOLTS'15) Elia, Halkidiki, Greece, 6-8 July 2015

doi: http://dx.doi.org/10.1109/IOLTS.2015.7229855

Abstract: Circuit aging causes a performance degradation and eventually a functional failure. It depends on the workload and the environmental condition of the system, which are hard to predict in early design phases resulting in pessimistic worst case design. Existing delay monitoring schemes measure the remaining slack of paths in the circuit, but cause a significant hardware penalty including global wiring. More importantly, the low sensitization ratio of long paths in applications may lead to a very low measurement frequency or even an unmonitored timing violation. In this work, we propose a delay monitor placement method by analyzing the topological circuit structure and sensitization of paths. The delay monitors are inserted at meticulously selected positions in the circuit, named observation points (OPs). This OP monitor placement method can reduce the number of inserted monitors by up to 98% compared to a placement at the end of long paths. The experimental validation shows the effectiveness of this aging indication, i.e. a monitor issues a timing alert always earlier than any imminent timing failure.

Preprint

Efficient Observation Point Selection for Aging Monitoring

Chang Liu, Michael A. Kochte and Hans-Joachim Wunderlich
ITI, University of Stuttgart, Pfaffenwaldring 47, D-70569, Stuttgart, Germany
Email: Chang.Liu@iti.uni-stuttgart.de, kochte@iti.uni-stuttgart.de, wu@informatik.uni-stuttgart.de

***Abstract*—Circuit aging causes a performance degradation and eventually a functional failure. It depends on the workload and the environmental condition of the system, which are hard to predict in early design phases resulting in pessimistic design. Existing delay monitoring schemes measure the remaining slack of paths in the circuit, but have a high hardware penalty including global wiring. More importantly, a low sensitization ratio of long paths in applications may lead to a very low measurement frequency or even unmonitored timing violations.**

In this work, we propose a delay monitor placement method by analyzing the topological circuit structure and sensitization of paths. The delay monitors are inserted at meticulously selected positions in the circuit, named observation points (OPs). This OP monitor placement method can reduce the number of inserted monitors by up to 98% compared to a placement at the end of long paths. The experimental validation shows the effectiveness of this aging indication, i.e. a monitor issues an alert always earlier than any imminent timing failure.

***Keywords*-Aging monitoring, delay monitoring, online test, concurrent test, stability checker, path selection**

I. Introduction

Circuit aging such as Negative-Bias Temperature Instability (NBTI), Hot-Carrier Injection (HCI), or electromigration (EM), causes a parameter shift in transistors or interconnects over the lifetime. As a result, the length of the critical or long paths may exceed the clock period and a timing failure occurs. A potential aging induced timing failure in safety- or security-critical applications like airplanes, automobiles, or power-plants could be life-threatening. Therefore, aging needs to be carefully monitored and appropriately handled. To avoid pessimistic static design margins, aging monitoring and adaptation approaches are often applied in fields. On-line built-in self-test techniques [1–4] and concurrent monitoring schemes have been proposed and can be applied to aging monitoring. The monitors sense the degradation effects in the circuit. Self-stressed monitors often include a stressed cell which is overcritically designed or overstressed, to guarantee that the stressed cell degrades faster than the circuit under actual operation [5, 6]. Tunable replica circuit [7] can be also categorized into such kind. The self-stressed monitors are often non-intrusive and with better hardware efficiency. However, their effectiveness bases on the assumption that the mission logic does not degrade faster than the stressed cell. Unfortunately this cannot be guaranteed since degradation strongly depends on transistor workload and working conditions (temperature or supply voltage). Due to large workload differences between applications [8, 9] self-stressed monitors have to be pessimistic and underestimate the real device lifetime. [10] synthesizes Representative Critical Reliability Paths (RCRPs) as a stand-alone circuit to estimate the aging degradation of the design in the field. This implementation improves the accuracy of aging estimation, but brings extra hardware cost and sampling control signals. Other effects, such as process variations, crosstalk and power supply noise, haven't been taken into account, which may also cause different delays in RCRPs and the functional circuit.

In-situ monitors measure a performance indicator directly in the functional part of the circuit. The Razor flip-flops in [11] are augmented with shadow latches synchronized by a delayed clock. By comparing the values in original registers and shadow latches, an aging induced timing failure can be detected. Canary logic [12] simplifies the Razor flip-flop design by eliminating the delayed clock. Delay detecting flip-flops [13–15] detect the degradation progress when a transition of the observed signal violates the predefined detection window (guard band) before the clock edge. Besides the detection window, the Scout FF in [16] also generates a tolerance window for late transitions after the clock trigger to predict as well as correct the failure. The low-cost transition detectors in [17] are based on transmission gates and allow to mask late transitions by shifting the clock. To reduce the hardware cost, the in-situ monitors can be placed at the end of the critical or long paths [13, 18]. In low power designs, the path lengths are equalized and many paths have lengths close to the critical path. Delay uncertainties caused by process variations, transistor workload or working conditions are hard to predict and compensate in the design phase. The critical path of the design may vary from chip to chip and over time. Thus, not only the nominal critical path but also the near critical ones have to be monitored. A non-enumerative technique [19] and a representative critical-path selection scheme [20] can reduce the number of paths necessary for observation.

SlackProbe [21] does not limit monitor placement to the end of paths, but also allows intermediate placement to save hardware cost. However, it does not take path sensitization into account when placing the monitors. Additionally, the control signals for the monitors may require global wiring or clock balancing schemes, increasing design complexity and possibly requiring to re-route the target design.

The monitor placement at the end of long paths implies that the path delay is only measured when the entire path is sensitized, which highly depends on the circuit structure, path length and input stimuli. Some applications hardly sensitize the whole circuit network [22], and it is possible that certain paths are only sensitized rarely during operation, causing large testing latencies and in the worst case unmonitored timing violations. Fig. 1 illustrates one of such

cases. A path through gates A, B and C, ABC for short, is a monitored long path. If ABC is seldom sensitized, the aged circuit may work fine without any alert or timing violation until ABC is activated. A transition along the degraded ABC violates not only the predefined timing margin but the clock as well. A timing violation occurs without indication.

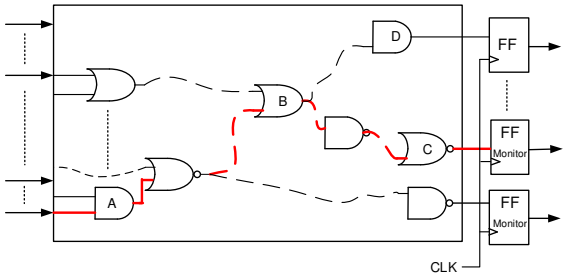

Figure 1. Limitation of the conventional monitor placement

To avoid such unmonitored timing violations and measure aging degradation at higher frequency, we propose a placement method for in-situ monitors. It analyzes the topological circuit structure, the path segment slack, and sensitization probability. Monitors are then inserted at meticulously selected positions in the circuit, named observation points (OPs), and measure the delay of path prefixes more frequently. Only few monitors are required to achieve a high coverage of target paths. The effectiveness of this approach is validated by simulation to show that a monitor activates always earlier than any imminent timing failure.

Section II gives an overview of the monitor placement scheme. The OP placement method is presented in Section III. The validation setup and experimental results are discussed in Sections IV and V.

II. Delay Monitor Placement Overview

A. Delay Detecting Flip-flop

A delay detecting flip-flop is a flip-flop extended with a delay monitor [13] (Fig. 2). The monitor can detect a delay increase along a sensitized circuit path w.r.t. the clock reference and a predefined timing margin called guard band.

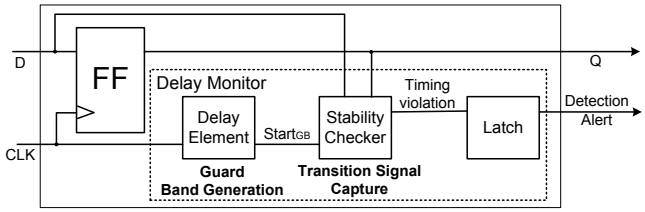

Figure 2. Structure of a delay detecting flip flop for degradation monitoring

A delay monitor consists of a delay element, stability checker and latch. The delay element shifts the clock to generate a reference signal ($Start_{\mathrm{GB}}$ in Fig. 3 (a)). The time period T_g between the rising edges of the clock (CLK) and $Start_{\mathrm{GB}}$ is the guard band, i.e. the time during which signal stability is checked. The design margin (Fig. 3 (a)) prevents a false activation due to process variation or early aging degradation. If in the nominal case the signal D_{Nominal} at a path endpoint reaches its stable value before the guard band, the path is not critical for operation. On the contrary, if the signal is unstable during the guard band (D_{Varied}), an alert is generated and stored in the latch.

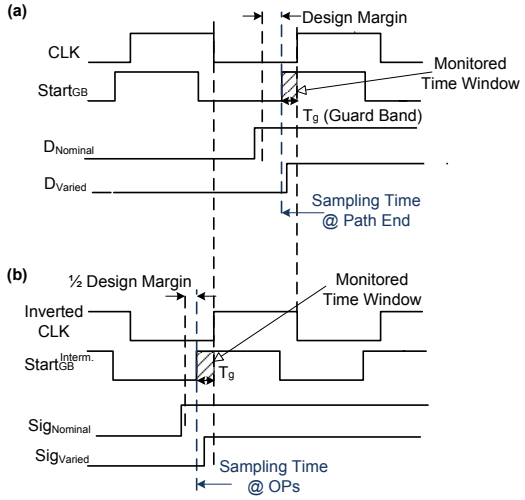

Figure 3. Signal waveform for a delay detecting flip-flop: (a) at path end; (b) at observation point (OP)

B. Delay Monitor Placement Approach

Our placement approach increases the measurement frequency of path delays and reduces the hardware costs compared to conventional endpoint placement. It selects observation points at intermediate nodes in the combinational circuit (e.g. OP in Fig. 4). Shorter paths or path prefixes are sensitized more often during operation, resulting in more frequent measurement of the path delay. Since path segments are typically shared by multiple paths, the selection of a subset of segments for monitoring allows to assess partial degradation of all or at least a high fraction of longer paths.

Positions closer to inputs are reached along shorter path prefixes, enlarging the sensitization probability and delay measurement frequency. These short prefixes are also impacted by aging. By monitoring such prefixes with lengths close to half the clock period, the inverted clock can be used as reference clock for the monitors to generate $Start_{\mathrm{GB}}^{\mathrm{Interm.}}$. The start of the monitored time window is then half the clock period minus the guard band (Fig. 3 (b)). Since the monitored path length is halved, the design margin shrinks proportionally to half of its original value.

To avoid complex global wiring and clock balancing issues in [21], the guard band and sampling time are unified for all selected observation points.

III. Selection of Observation Points (OP)

A. Terminology and Problem Statement

Let the *topological depth of a gate* g be the length of the longest path segment from an input to g. The set of *target paths* comprises the critical and near-critical long paths in the circuit, obtained e.g. by static timing analysis (STA). The

target outputs are the primary and pseudo-primary outputs at the end of target paths. An *OP covers a path* if the delay of the path initial segment, which can be directly measured by a monitor inserted at OP, exceeds a predefined threshold (Section III-B). The *target path coverage* is the percentage of target paths covered by OPs. The *OP candidates* are possible monitor insertion locations, limited by timing constrains to reduce the search space of OP selection (Section III-C). The *OP slack* or *OP candidate slack* is the time difference between the start of the monitored time window (half the clock period minus guard band) and the topological depth of the OP or OP candidate. In the nominal case, the OP slack is always larger than or equal to half the design margin (Fig. 5 (b)). The *OP upper bound* (OP_{UB}) bounds the topological depth of the OP candidates. As mentioned in Section II, half the clock period is chosen as sampling reference and every transition violating the sampling time will trigger a detecting alert. To prevent monitor false activation, OP_{UB} is set as half the clock period minus guard band and half the design margin. A *path prefix* is the prefix of a path starting from its primary or pseudo-primary input to the gate with maximum topological depth less than OP_{UB}.

Selecting OPs among the candidates to cover the target paths is a set-covering problem with two objectives: the number of OPs should be minimized for low hardware cost, and the selected OPs should have high sensitization probability during operation to ensure frequent measurement of covered target paths. The search space is initially narrowed down by identifying the OP candidate range (Section III-C). Then the heuristic OP selection algorithm (Section III-D) is used to find a small set of OPs among the candidates.

B. Target Path Coverage of Observation Points

The principle of our monitoring approach is to assess the degradation by measuring the delay of target path prefixes. Many target paths have branches close to OP_{UB}. The branched prefixes, e.g. segments ABF and ABC in Fig. 4, share most of the path elements. The large common path prefix AB implies a high correlation between the lengths of the branches ABF and ABC, allowing to estimate the length of all branches by observing one of them. If the length of the common prefix (as percentage of target prefix length) exceeds a given threshold (e.g. $length_{\text{AB}}/length_{\text{ABF}} > 70\%$), we say the target path ABG is partially covered by the OP (blue point at gate C), although the corresponding output O_1 is unreachable from the OP at C.

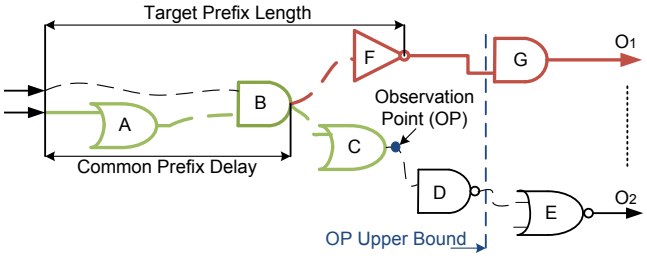

Figure 4. OP covering a target path prefix

C. Observation Point Candidate Range

In principle, all gates with a depth less than OP_{UB} are potential OP candidates. To reduce the search space for OP selection, a candidate range is introduced that limits the considered candidates. It is computed by balancing the monitoring quality and overhead. The quality depends on the target path coverage and the OP candidate slacks.

If a candidate's depth is close to OP_{UB} (e.g. OP_1 in Fig. 5 (a)), the slack of the candidate will be close to the design margin (light gray area shown in Fig. 5 (b)). During operation, OP_1 slack gradually reduces due to degradation and may eventually cause a timing alert by the monitor. However, it may be that the candidates close to OP_{UB} cannot cover all target paths. For high target path coverage, the gates with larger slacks are also considered as candidates (e.g. OP_2). To maintain a similar margin for degradation, the time difference between the OP slack and half the design margin, named delay matching region, is compensated by inserting delay elements (e.g. inverters).

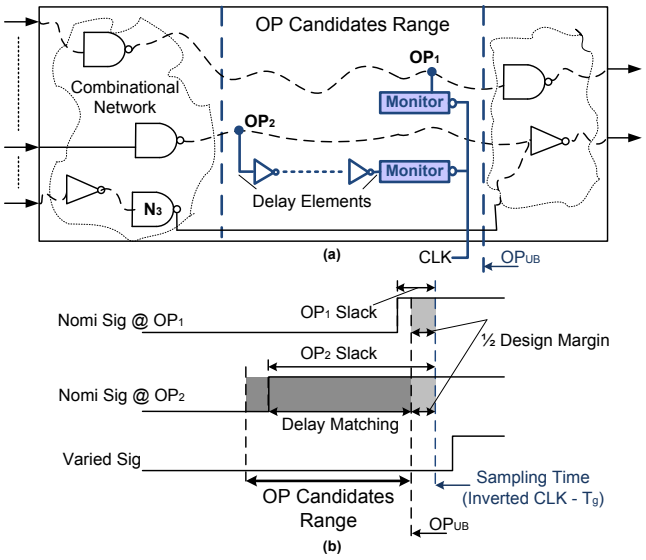

Figure 5. OP candidate range identification: (a) topological position in the circuit netlist; (b) signal waveform

The required number of inverters can be calculated based on the OP slack in either nominal or degraded case. Since the degradation of prefixes and paths depends on the workload and operation condition, the prefix segments can degrade at varying speed with the corresponding paths. Usually the general application scenarios and average operation condition are already known at design time. They can be utilized for aging analysis to adjust the OP delay matching and further improve the monitoring accuracy. The dark gray area in Fig. 5 (b) illustrates the topological range of OP candidates ($TopoDelay_{\text{cand}}$):

$$OP_{\text{UB}} - MD_{\text{max}} < TopoDelay_{\text{cand}} < OP_{\text{UB}} \qquad (1)$$

MD_{max} is the maximal matching delay of the inverter chain. The OP candidate range should satisfy the following criteria:

- The OP candidates should provide a high target path coverage to avoid insertion of extra delay monitors at the end of uncovered target paths.
- The hardware overhead caused by delay matching should be minimal.
- The path prefixes covered by OP candidates should contain enough degradation information for aging monitoring, i.e. the prefix length should be long enough to ensure high timing correlation to the entire target path.

Due to the second and third criterion, a full target path coverage can't always be achieved by simply increasing the matching delay (length of the inverter chain). For instance, as shown in Fig. 5 (a), for the path at the bottom of the figure, N_3 is the gate closest to OP_{UB}. However, if the path prefix length (propagation delay from input to N_3) is not large enough, the delay degradation from the inverter chain may dominate the overall degradation of the path prefix. In this case, to keep the monitoring quality, sensing the entire path length would be a better solution than measuring only the prefix segment. As shown later in the experiment results, such uncovered paths are very rare.

D. OP Selection Algorithm

The OP selection problem can be formulated as follow: for a computed set of OP candidates C (Section III-C), identify a minimal subset $V \subseteq C$ (i.e. OPs) such that the delay degradation of the target paths can be observed by monitors placed at V with a high measurement frequency. The measurement frequency is quantified by the OP sensitization ratio, generated by logic simulation of functional or random patterns. It is calculated as the number of cycles in which the signal at an OP toggles divided by the total number of simulation cycles. For each candidate, a merit factor μ is computed as the product of the number of covered paths and the sensitization ratio. The candidates are selected iteratively as OPs in order of decreasing merit μ. To reduce the coverage overlap, the paths covered by fewer candidates are considered with higher priority.

Before OP selection, the netlist and target paths $P := \{p_1, p_2, p_3, \ldots, p_N\}$ are used to compute the OP candidates $C := \{c_1, c_2, c_3, \ldots, c_M\}$. Let $sr(c_i)$ be the sensitization ratio of candidate c_i. The target path coverage is calculated w.r.t. the entire candidate group. As discussed in Section III-C, uncovered paths ($U \subseteq P$) are removed from the target group ($P := P \setminus U$), and delay monitors will be integrated into flip-flops at the end of uncovered paths.

Then, OP selection iterates as follows:

- Step 1: Analyze the path and OP candidate relation
  - For every target path p_i, the OP candidates covering the path are identified, denoted as $SC_i \subseteq C$. Let $\mid SC_i \mid$ be the size of SC_i.
  - For every candidate c_i, the set of covered paths is defined as $SP_i \subseteq P$. $\mid SP_i \mid$ is the size of SP_i.
- Step 2: To reduce the path coverage overlap and obtain an optimal OP set, the hard to be covered paths (i.e. paths covered by a minimal number of OP candidates) are considered prior in OP selection: select p_s, when $\mid SC_s \mid = \mathtt{MIN}_{i=1}^{N} \{\mid SC_i \mid\}$. SC_s is the candidate set covering path p_s. This set is irreplaceable by other candidates for a high target path coverage.
- Step 3: Select the candidate c_s in SC_s with maximal merit factor $\mu_s = \mathtt{MAX}_{i=1}^{|SC_s|} \{\mid SP_i \mid \times sr(c_i)\}$ as OP. This OP c_s covers not only the hard to cover path p_s, but has also a large potential to cover other target paths with high measurement frequency.
- Step 4: Remove the selected OP and covered target paths: $C := C \setminus c_s$, $P := P \setminus SP_s$. If $P \neq \emptyset$, repeat from step 1.

IV. OP Effectiveness Validation

The effectiveness of the selected OPs are evaluated from two aspects: (1) Failure predictability: a timing alert is generated at OPs before any actual timing failure, and (2) Prediction validity: every monitor alert refers to an imminent failure, i.e. no false positive alerts are generated.

In the validation, the transistor stress is estimated by logic simulation of functional or random input stimuli. The nominal standard delay file (SDF) and the transistor stress are input to an aging model to create the SDF for degraded gates. Using the degraded delay information, timing simulation is done for an input pattern set. If a transition at an OP violates the inverted clock, an aging alert is issued. A transition at a target output exceeding the clock period indicates a timing failure.

Different degraded timing profiles are generated per circuit for different system operation times. For each degraded SDF, timing simulation is repeated and the transitions at OPs and outputs are analyzed. The time of the first alert activation and the first timing failure at outputs is recorded. If the first alert occurs earlier than any failure, the property of failure predictability holds.

To show the prediction validity, the remaining timing margin of the nominal critical path is analyzed when the first aging alert occurs at an OP.

V. Results of the Experiments

We evaluate the approach on ISCAS'89 and NXP benchmarks. The Nangate 45 nm open cell library is used for synthesis and timing analysis. The design margin is set to 10% of the critical path length (cpl), i.e. $clock = 1.10 \cdot cpl$.

A. Observation Point Selection Results

The results of OP selection are listed in Table I. The number of gates in the target design is shown in column 2. The longest 100 paths per output with a length from $70\% \cdot cpl$ to the critical path length through this output are selected as the target path group using STA. The number of target paths is presented in column 3. The number of (primary and pseudo-primary) target outputs is in column 6.

The maximal matching delay MD_{max} in Eq. (1) is set to the delay of six inverters. The candidates covering no target paths are removed from the group. The number of OP candidates after elimination is presented in column 4.

The target path coverage of the OP candidates is shown in column 5. A high target path coverage is achieved by the candidate range defined above and this coverage is maintained during OP selection.

The prefix covering threshold (Section III-B) is set to 70%, i.e. for every target path at least 70% of its prefix delay is directly measured by the corresponding OP. The number of OPs and the number of sensors inserted at path ends are listed in columns 7 and 8 respectively. The monitor number reduction $red. := (\#tar\_out - \#OP - \#end)/\#tar\_out$ is provided in the last column of the table.

Table I. OP selection results

circuit	#gates	#paths	#cand.	cov.	#tar_out	#OP	#end	red.
s9234	1764	2284	33	0.99	55	13	9	0.60
s13207	2865	346	49	0.94	18	16	1	0.06
s15850	3320	4937	49	1.00	78	7	0	0.91
s35932	11168	6833	472	1.00	320	184	0	0.43
s38417	9796	12402	104	0.99	197	28	2	0.85
s38584	12183	622	21	0.95	47	11	6	0.64
p35k	23294	6207	2	1.00	71	2	0	0.97
p45k	25406	3727	37	1.00	54	4	0	0.93
p78k	70495	132545	2227	1.00	1417	306	0	0.78
p81k	82265	6268	112	1.00	69	3	0	0.96
p89k	58726	34184	62	1.00	352	16	0	0.95
p100k	60767	11151	271	1.00	116	36	0	0.69
p141k	107655	32485	186	1.00	333	8	0	0.98

Compared to conventional endpoint placement, the proposed method reduces the number of monitors significantly. The reduction of inserted monitors ranges from 6% (s13207) up to 98% (p141k). For larger circuits, the reduction improves. On average, the reduction is 58% for the ISCAS'89 circuits and 89% for the NXP circuits.

According to the transistor schematic [13], the stability checkers consume dynamic power only when a late transition violates the guard band. The signal transition frequency, i.e. the sensor measurement frequency, won't influence the power consumption unless an aging alert is triggered. Consequently, the power computation of the sensors reduces proportionally to the reduction of sensors.

The experiment is processed on a Xeon server with 2.67GHz CPUs and 80 GB memory. The run-time of the OP selection algorithm is ca. 6 minutes for p141k.

B. Results of OP Effectiveness Validation

The goal of the validation experiment of Section IV is to investigate if an earlier prediction can be achieved by more frequent path prefix measurement in the proposed method. Since the number of random patterns needed to sensitize a path grows exponentially with the path length, for deep circuits long target paths will be sensitized only with a very low probability. In the validation experiments, the limited number of random patterns is insufficient to sensitize a significant fraction of paths through the OPs and target paths. For that reason, path delay fault ATPG patterns for the target paths are generated and applied in the OP validation flow. Due to the ATPG abort limit, some sensitizable target paths may not be activated during validation.

Additionally, 10240 random patterns are simulated to calculate the workload-induced stress of each transistor. The workload is assumed to be constant for the circuit lifetime. NBTI is considered as the dominant mechanism of aging degradation. Similar to [19], the NBTI model published in [23] is used here for aging simulation. According to the aging model, the gate delay increase (Δt) is computed as: $\Delta t = A{\cdot}(\alpha{\cdot}t)^n{\cdot}t_0$. The stress probability α is the probability that a PMOS transistor is under stress in one cycle. t refers to the circuit total operation time. $n = 1/6$ is a characteristic constant of the NBTI effect. t_0 denotes the nominal pin to pin delay of the gate. A is a constant parameter and is adjusted so that the delay degradation is 10% over 5 years under 50% stress probability. Different operation times from 0 to 10 years with the resolution of a quarter year (0.25, 0.5 ... 9.75, 10 years) are applied to the aging model to generate the degraded gate delay over the operation time. Later, the above mentioned path delay fault patterns are used in the timing simulation. The timing simulation is repeated with the degraded delays of the different operation times until 10 years. The results are listed in Tables II and III. Dashes in the table indicate that aging alerts were not issued or timing failures not caused by the applied input stimuli during 10 years of operation. For circuit s13207, no aging alert or timing failure was observered at all.

Since the inverter chain for OP delay matching can be implemented based on nominal or degraded timing (Section III-C), the two cases are investigated separately:

1) Delay Matching based on Nominal Timing Profile: In Table II, column 2 to 4 provide the operation time (in years) until the occurrence of an event. The event can refer to the first observed aging alert activation (OP_act), the first observed timing failure ($failure$, both based on timing simulation), or the first clock violation by the critical path (cpl_vio, computed by STA). For most of circuits, the first aging alert occurs earlier than timing failures. However, for circuits s15850 and p35k, the monitored prefixes degrade slower than the entire paths. This can be avoided when the workload and operating condition of the circuit are applied for delay matching (cf. Section V-B2). For some circuits (e.g. s38417, s38584), a timing failure occurs before the degraded critical path violates the clock period. This is because the critical path changes during degradation.

The last two columns display the remaining time margin of the nominal critical path when the first monitor alert activates. The absolute time margins in femtoseconds and the relative values as the percentage of the clock period are listed. The remaining margin is less than 1% in average, illustrating the degradation process in the circuits.

A timing violation of the nominal critical path indicates a potential aging failure, but it may not be activated by the input stimuli. For p141k, for instance, the critical path violates the timing after 7.5 years. This degradation could be undetectable if the monitors are placed at path endpoints and the entire path is not sensitized by the input stimuli. As shown, no timing failure has been observed during 10 years of operation (dash in last row of the table). By measuring the path prefix, this potential failure is indicated in the 8.25th

year. Still, the OP activation occurs later than the potential failure (marked red) since the nominal timing profile is used.

Table II. OP validation (delay matching based on nominal profile)

circuit	OP_act	failure	cpl_vio	re_margin (fs)	%clk
s9234	5.25	8.50	8.50	6681	0.70%
s13207	–	–	–	–	–
s15850	8.00	7.00	9.50	4613	0.22%
s35932	5.50	8.25	8.25	2631	0.59%
s38417	3.75	5.50	–	19469	1.56%
s38584	3.00	5.50	7.00	14007	1.16%
p35k	6.75	2.50	8.50	11085	0.32%
p45k	7.75	9.75	9.75	8417	0.33%
p78k	6.00	8.00	8.00	7174	0.39%
p81k	7.00	8.75	8.75	17646	0.29%
p89k	6.00	8.75	7.25	7430	0.23%
p100k	7.50	–	9.75	10648	0.35%
p141k	8.25	–	7.50	-5292	-0.18%

2) Delay Matching based on Degraded Timing Profile: By using the degraded timing profile for delay matching, not only the problems in red can be solved (bold font in Table III), but for rest of the circuits also the time of the first alert is less pessimistic. The year of first timing failure occurrence and the clock violation time of the nominal critical path (values in column 3 and 4) are identical in both tables. The results in the last two columns only differ from the previous table if the time of the first timing alert changes.

Table III. OP validation (delay matching based on degraded profile)

circuit	OP_act	failure	cpl_vio	re_margin (fs)	%clk
s9234	5.25	8.50	8.50	6681	0.70%
s13207	–	–	–	–	–
s15850	**5.75**	7.00	9.50	14600	0.69%
s35932	5.50	8.25	8.25	2631	0.59%
s38417	3.75	5.50	–	19469	1.56%
s38584	4.00	5.50	7.00	9302	0.77%
p35k	**1.75**	2.50	8.50	72437	2.09%
p45k	7.75	9.75	9.75	8417	0.33%
p78k	6.00	8.00	8.00	7174	0.39%
p81k	8.25	8.75	8.75	2757	0.05%
p89k	6.50	8.75	7.25	3662	0.12%
p100k	8.50	–	9.75	5102	0.17%
p141k	**5.50**	–	7.50	**12932**	**0.43%**

VI. Conclusion

This paper presents a new delay monitor placement method, reducing both the measurement latency and monitor overhead. The method utilizes the inverted clock as the unified monitor sampling time, therefore avoiding global wiring and clock balancing problems induced by monitor insertion at arbitrary nets. The proposed algorithm takes target path branches in the circuit structure and gate sensitization probability during operation into account. The results show that our approach reduces the number of required monitors by up to 98%. Due to the high measurement frequency at OPs, unmonitored timing violations can be avoided. In the experimental validation the comparison of alerts and failure occurrences shows that this cost-efficient placement effectively indicates immiment timing failures.

Acknowledgements

This work was supported by the German Research Foundation (DFG) under grant WU 245/13-1 (RMBIST).

References

[1] Y. Li, S. Makar, and S. Mitra, "CASP: Concurrent Autonomous Chip Self-Test Using Stored Test Patterns," in *Design, Automation and Test in Europe (DATE)*, March 2008, pp. 885–890.

[2] M. A. Kochte, C. G. Zoellin, and H.-J. Wunderlich, "Concurrent self-test with partially specified patterns for low test latency and overhead," in *Proc. IEEE European Test Symposium (ETS)*, 2009, pp. 53–58.

[3] Y. Sato, S. Kajihara *et al.*, "A circuit failure prediction mechanism (DART) for high field reliability," in *Proc. IEEE International Conference on ASIC (ASICON)*, Oct 2009, pp. 581–584.

[4] S. Hellebrand, T. Indlekofer *et al.*, "FAST-BIST: Faster-than-At-Speed BIST Targeting Hidden Delay Defects," in *Proc. IEEE International Test Conference (ITC)*, 2014, pp. 1–8.

[5] R. Carlsten, J. Ralston-Good, and D. Goodman, "An Approach to Detect Negative Bias Temperature Instability (NBTI) in Ultra-Deep Submicron Technologies," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2007, pp. 1257 –1260.

[6] T.-H. Kim, R. Persaud, and C. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 874–880, April 2008.

[7] J. Tschanz, K. Bowman *et al.*, "Tunable Replica Circuits and Adaptive Voltage-Frequency Techniques for Dynamic Voltage, Temperature, and Aging Variation Tolerance," in *Proc. 2009 Symposium on VLSI Circuits*, June 2009, pp. 112–113.

[8] V. Chandra, "Monitoring Reliability in Embedded Processors - A Multi-layer View," in *Proc. ACM/IEEE Design Automation Conference (DAC)*, 2014, pp. 46:1–46:6.

[9] R. Baranowski, F. Firouzi *et al.*, "On-Line Prediction of NBTI-induced Aging Rates," in *Proc. Design, Automation & Test in Europe Conference (DATE)*. EDA Consortium, 2015, pp. 589–592.

[10] S. Wang, J. Chen, and M. Tehranipoor, "Representative Critical Reliability Paths for Low-Cost and Accurate On-Chip Aging Evaluation," in *IEEE/ACM Int'l Conf. on Computer-Aided Design (ICCAD)*, Nov 2012, pp. 736–741.

[11] S. Das, C. Tokunaga *et al.*, "RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan 2009.

[12] T. Sato and Y. Kunitake, "A Simple Flip-Flop Circuit for Typical-Case Designs for DFM," in *Proc. International Symposium on Quality Electronic Design (ISQED)*, March 2007, pp. 539–544.

[13] M. Agarwal, V. Balakrishnan *et al.*, "Optimized Circuit Failure Prediction for Aging: Practicality and Promise," in *Proc. IEEE International Test Conference (ITC)*, Oct. 2008, pp. 1–10.

[14] H. Dadgour and K. Banerjee, "Aging-resilient design of pipelined architectures using novel detection and correction circuits," in *Proc. Design, Autom. and Test in Europe Conf. (DATE)*, 2010, pp. 244–249.

[15] J. Vazquez, V. Champac *et al.*, "Programmable aging sensor for automotive safety-critical applications," in *Proc. Design, Automation Test in Europe Conf. (DATE)*, march 2010, pp. 618–621.

[16] J. Semiao, D. Saraiva *et al.*, "Performance Sensor for Tolerance and Predictive Detection of Delay-Faults," in *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, Oct 2014, pp. 110–115.

[17] M. Omaña, D. Rossi *et al.*, "Low Cost NBTI Degradation Detection and Masking Approaches," *IEEE Transactions on Computers*, vol. 62, no. 3, pp. 496–509, March 2013.

[18] W. Wang, Z. Wei *et al.*, "An Efficient Method to Identify Critical Gates under Circuit Aging," in *Proc. IEEE/ACM Int'l Conf. on Computer-Aided Design (ICCAD)*, Nov. 2007, pp. 735–740.

[19] A. Baba and S. Mitra, "Testing for transistor aging," in *Proceedings of 27th IEEE VLSI Test Symposium (VTS)*, May 2009, pp. 215–220.

[20] F. Firouzi, F. Ye *et al.*, "Representative Critical-Path Selection for Aging-Induced Delay Monitoring," in *Proc. IEEE International Test Conference (ITC)*, Sept 2013, pp. 1–10.

[21] L. Lai, V. Chandra *et al.*, "SlackProbe: A low overhead in situ on-line timing slack monitoring methodology," in *Design, Automation Test in Europe Conf. (DATE)*, March 2013, pp. 282–287.

[22] H. Yalcin, R. Palermo *et al.*, "An advanced timing characterization method using mode dependency," in *Proc. ACM/IEEE Design Automation Conference (DAC)*, 2001, pp. 657–660.

[23] S. Bhardwaj, W. Wang *et al.*, "Predictive Modeling of the NBTI Effect for Reliable Design," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2006, pp. 189–192.