Chip Health Tracking Using Dynamic In-Situ Delay Monitoring

Hadi Ahmadi Balef
Eindhoven University of Technology
Eindhoven, The Netherlands
h.ahmadi.balef@tue.nl

Kees Goossens
Eindhoven University of Technology
Eindhoven, The Netherlands
k.g.w.goossens@tue.nl

José Pineda de Gyvez
NXP Semiconductors
Eindhoven University of Technology
Eindhoven, The Netherlands
j.pineda.de.gyvez@tue.nl

***Abstract*— Tracking the gradual effect of silicon aging requires fine-grain slack monitoring. Conventional slack monitoring techniques intend to measure worst-case static slack, *i.e.* the slack of longest timing path. In sharp contrast to the conventional techniques, we propose a novel technique that is based on dynamic excitation of in-situ delay monitors, *i.e.* dynamic excitation of the timing paths that are monitored. As the delays degrade, the path delays increase and the monitors are excited more frequently. With the proposed technique, a fine-grained signature of the delay degradation is extracted from the excitation rate of monitors.**

Keywords—delay testing, in-situ monitoring, silicon aging, reliability

I. INTRODUCTION

In the nanoscale era, circuit components have become more unpredictable, making design of reliable electronic systems more challenging. On the other hand, the demand for power reduction (*e.g.* for battery powered applications) is booming and this adds to the design uncertainties. That is because by lowering supply voltage, as an effective technique for power reduction, the uncertainty effects are amplified [1]. Over-engineering is a traditional means for reliable system design. However, in the presence of dramatic uncertainty effects, design efficiency is jeopardized with over-engineering. To increase design efficiency, chip health tracking systems play a key role in contemporary circuit design paradigms. Those techniques enable a feedback loop to tune the operating point of circuit according to the actual chip health status. Chip health tracking systems are not only intended to solve the overdesign cost problem, but also they are vital for safety critical applications. In those applications, reliable functionality of system must be guaranteed, thereby mandating continuous monitoring of chip health status.

The state-of-the-art chip health tracking techniques for aging are either based on monitoring the stress factors [2], sensing the parameters of a transistor under stress [3], monitoring the delay of a replica path or a ring-oscillator [4], in-situ delay monitoring [5], or delay testing [6]. The main advantage of in-situ delay monitoring and delay testing is that they enable direct sensing of critical path slack. With delay testing, the test time is long while the coverage is limited, because exciting all critical paths with the test approach is very expensive. On the other hand, in-situ delay monitoring relies on the actual workload of the circuit for path excitation to detect delay degradation. Therefore, the monitoring task is performed at runtime with minimal functional intrusion and hardware overhead. Traditional in-situ monitoring techniques insert the monitors at the end point of the timing paths [7]. However, due to logic masking, the Monitor Excitation Rate (MER) is much less when the monitors are inserted at the end points. MER is defined as the average number of cycles in which a monitor is excited. In [5], the monitors are inserted at intermediate points along timing paths to increase MER and to reduce the number of monitors. With higher MER, the observability of the monitors (defined as the average number of monitored nets per cycle [5]) increases. In [8], the insertion points are selected deeper inside timing paths to increase MER and the cost of efficiency even further. The conventional in-situ monitoring techniques only check whether or not the slack margins are tight (*i.e.* binary output). To get a fine-grain measurement of delay, TDC is employed in [9]. Although TDC is a reasonable choice for the replica path based on ring-oscillator monitoring techniques, inserting a TDC at each critical path increases hardware overhead substantially.

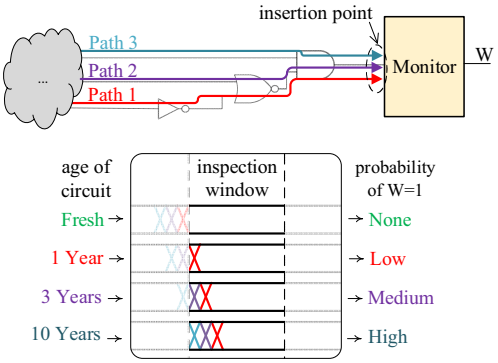

Figure 1 - The proposed idea.

The state-of-the-art static slack monitoring techniques do not use the underlying hardware in the most efficient way. This work proposes an innovative technique of in-situ delay monitoring that instead of measuring the static slack, it enables fine-grain monitoring of delay degradation based on dynamic monitor excitation. With our technique, MER is used as an indication of delay degradation. The proposed idea is illustrated in Figure 1. Suppose three paths with different delays are ending into the monitor, as illustrated in the figure. When the monitor is excited, its output W becomes a logic one. As path delays increase, more paths can cause transitions during the inspection window. Therefore, the excitation rate of the monitor increases when delays degrade, *i.e.*, the probability that the monitor's output is W=1 increases. Hence, there is a strong positive correlation between delay degradation and MER. We employ this relation to generate a fine-grained signature of chip health status with less hardware overhead.

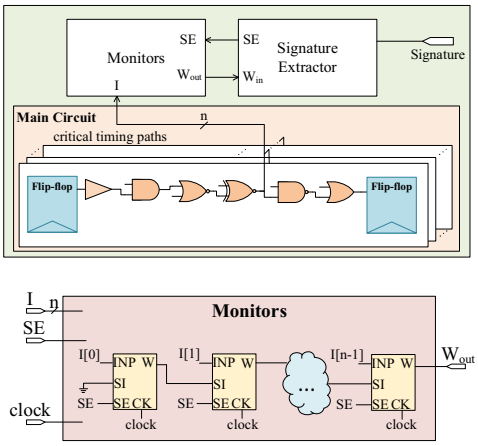

Figure 3 - The proposed chip health tracking system and the Monitors block which consists of the monitors connected in a scan chain.

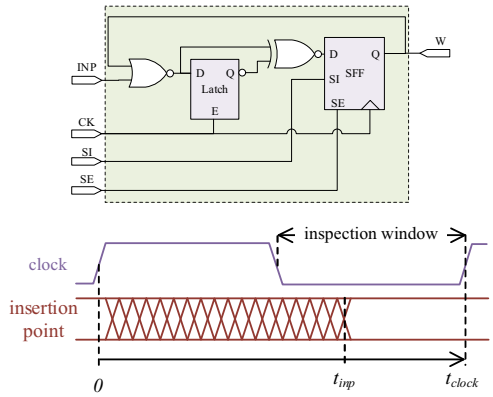

Figure 2 – The design of monitor and activity of monitor insertion point.

II. The Proposed Chip Health Tracking System

The architecture of the proposed chip health tracking system is shown in Figure 2. This system is made up of the monitors and the signature extractor. As shown in the figure, the monitors are connected to internal circuit nodes along timing critical paths. Delay degradation is monitored by sensing the arrival time of data to the insertion points of monitors. The monitoring information is then scanned out periodically to be analyzed with the signature extractor block. Accordingly, the monitors are either in the monitoring mode or in the scan mode and this is determined by SE, *i.e.*, when SE is high (low), monitors are in scan (monitoring) mode. At the end of each scan phase, all monitors are reset to zero for the next monitoring phase. In this way, a time sampling of monitoring excitation is implemented. The signature extractor continuously analyses the output of monitors and generates the reliability signature based on MER.

The proposed system outputs a fine-grained slack measurement if the sensitivity of MER to delay degradation is significant. To achieve this, the monitors are inserted selectively at the intermediate points along timing paths so that the logic masking effect is reduced. Assuming that the design margins are tight, the sensitivity of MER to delay degradation is significant. As shown in Figure 2, the monitors are connected in a scan chain. The output of the last monitor in the chain is connected to W_{out} and the input of the first monitor in the chain is tied low. During the scan phase, the stored values in the monitors are scanned out and logic zero propagates to all scan flip-flops of monitors from the input of the first monitor. Therefore, all monitors are reset to zero at the end of scan phase. Note that the monitoring data stream has no intrusion to the main data stream of circuit. Furthermore, during the scan mode, there is no unnecessary toggling in the main circuit. Hence, the chip health tracking is performed in parallel to the main functionality at the nominal speed.

III. The Monitors

We take the design and the insertion flow of [8] for in-situ monitoring, and improve their work to be used in our chip health tracking system. The design of the monitor is shown in Figure 3. As explained before, the monitors are either in the monitoring mode (SE=0) or in the scan mode (SE=1). In the monitoring mode, the monitor detects late data arrival with a latch and an XOR gate. If the last transition at the insertion point occurs in the second half of clock cycle, the output of the XOR gate becomes logic one. If a late arrival is detected, logic one is stored in the output scan flip-flop. With a NOR gate, logic one stays in the scan flip-flop as long as the monitor is in the monitoring mode. As discussed before, in-situ delay monitoring relies on path excitation. If the monitored path is not excited, the monitor fails to capture the degradation effect. Therefore, it is essential to monitor more than one path to increase the sensitivity of MER to delay degradation. Furthermore, in the presence of variation effects, the critical path of the circuit can change [10]. The paths with lower slack are more critical and have more priority for monitoring. All paths whose slack is less than a specific value are monitored and the monitors are inserted at intermediate points of those paths. The input activity of the monitor is illustrated in Figure 3. The inspection window opens from $t=0.5\times t_{clock}$, and it closes at $t=t_{clock}$, where t_{clock} is the clock period. Since multiple paths end into the monitor, activity at the insertion point can start from the beginning of clock cycle, depending on the delay of shortest path to the insertion point. The last transition during each clock cycle can happen at $t=t_{inp}$, where t_{inp} is equal to the maximum delay to the insertion point. Therefore, t_{inp} affects MER by affecting the chance of transition during the inspection window. The insertion points are identified based on t_{inp} for a given set of paths.

IV. The Signature Extractor

As introduced in section II, the signature extractor analyzes the monitoring data to generate a reliability signature based on MER. To get MER, the monitors are periodically reset to zero after reading the monitoring data. In this way, a time-sampling of monitor excitation is implemented. In Figure 4, the signature extractor is illustrated. SE controls the sampling period and it is generated with a counter which up-counts to a specified *period* and starts from zero again. When the counter value is less than n_{mon}, SE becomes a logic one to scan out all monitoring data. As discussed before, the monitors are reset to zero after this by propagating a logic zero in the scan chain. Then, SE becomes logic zero for (*period*-n_{mon}) cycles and during this time the monitors capture late data arrivals. Intuitively, when circuit

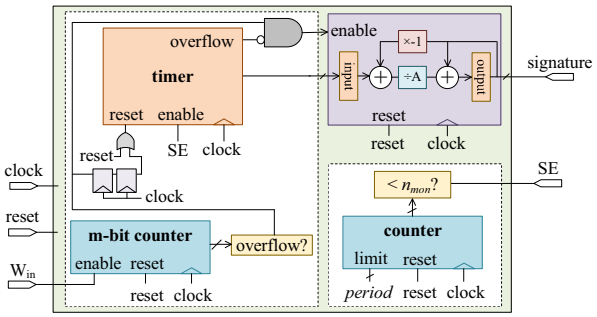

Figure 5 - The design of signature extractor.

delays increase, more monitors are excited, and each monitor is excited more frequently. To get an indication of MER, the overall number of excitations is counted continuously in the signature extractor, with an m-bit counter that is enabled with W_{in}. The counter up-counts until it reaches to its limit 2^{m-1} and starts from zero again. Hence, with higher MER the m-bit counter overflows faster. The overflow of m-bit counter is flagged with a signal that is high for one clock cycle. Therefore, more frequent overflows indicates that the delay degradation effect is more severe. To measure this frequency, a timer is used to get the number of cycles between consecutive overflows of m-bit counter. The output of the timer is sampled when the m-bit counter overflows and the timer is reset to zero afterwards. The final signature is obtained by calculating a weighted average of the sampled timer value as

$$S[i] = \frac{(A-1) \times S[i-1] + T[i]}{A}, \quad (1)$$

where i is the sample number, A is the weight factor, S is the signature, and T is the output of the timer. Since m-bit counter is only enabled during scan mode (W_{in} can be logic one only in this mode), the timer is disabled during monitoring mode by connecting the enable port of the timer to SE. As shown in Figure 4, the overflow signal coming from the m-bit counter is combined with overflow signal of the timer, to get the enable signal for sampling. That is because if the m-bit counter does not overflow or it overflows very rarely, the timer overflows. In this case, the timer value is not valid and it should not be sampled. Therefore, the delay degradation when the signature generation starts is determined according to the bit-width of the timer. If the timer's bit-width is low, then it overflows faster and the signature generation only starts if the m-bit counter overflows more frequently, *i.e.* when the delay degradation is more. On the other hand, if the timer's bit-width is high, it enables measuring the less frequent overflows of the m-bit counter, thereby the signature is generated for lower delay degradations. Therefore, the bit-width of the timer determines the threshold of the signature generation.

V. Experimental Results

The proposed chip health tracking system is evaluated with an ARM Cortex M0 processor in 40nm technology. The target speed for the design is 200MHz. Cadence tools for front-end and back-end design, namely Genus and Innovus, are used for implementation. Timing analysis is performed with Cadence Tempus, and netlist simulation is performed with Cadence Incisive. The design corners are slow and fast for setup and hold analysis, respectively. To have a realistic sense of design

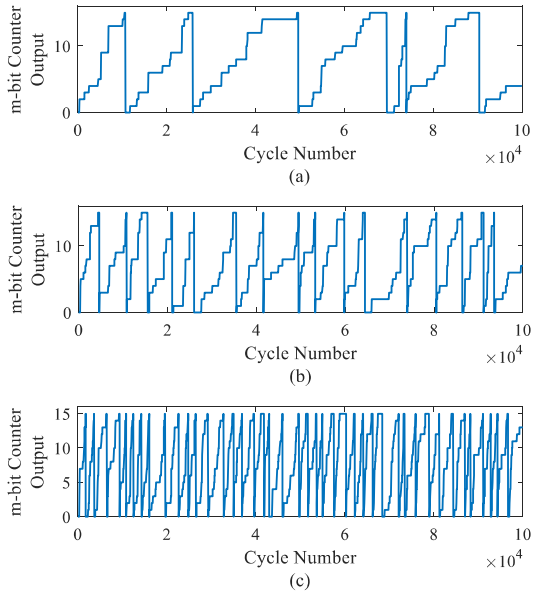

Figure 4 – The output of the m-bit counter before overflow detection considering three delay scaling factors (a) *1.4X*, (b) *1.5X*, and (c) *1.6X*.

margins, the typical corner is considered for netlist simulation with timing. In the following subsections, the proposed aging simulation framework is explained first, and then the proposed system is evaluated in different ways. We used the NBTI model from [11], and assumed that the threshold voltage of transistors degrades by 10% after 10 years of constant stress. To apply this degradation to the timing analysis, derating factors are specified for each of the gate delays according to the obtained delay degradation. For netlist simulation of the aged circuit, SDF file is dumped from the timing analysis tool. Furthermore, delay scaling factor is applied during timing annotation from the SDF file.

We performed a netlist simulation with annotated timings to evaluate the functionality of our chip health tracking system. The simulation testbench runs an FFT application on the processor for 100K cycles. 64 monitors are added to the design with t_{inp}=0.6×t_{clock}. The period of SE is 128 cycles, which means the monitors are in the monitoring mode for 64 cycles, as explained in the previous section. The output of m-bit counter is shown in Figure 5. For this simulation, three delay scaling factors are applied: *1.4X*, *1.5X*, and *1.6X*. Note that based on our experiments, the delays can be scaled up to *~1.75X* without having any timing error in the main design. As can be observed, with higher delay scaling factor, the frequency of overflow increases. That is because MER is higher and this results in more countings with m-bit counter. In this experiment, the bit-width of the m-bit counter is m=4. With a higher (lower) m, the frequency of overflowing decreases (increases).

In Figure 6, the output signature is shown considering different delay scaling factors. The bit-width of timer is 11 and the weight factor in (1) is A=8. As can be observed, when the delay scaling factor is *1.5X*, the signature is zero, *i.e.* the signature is not generated. That is because up to this delay

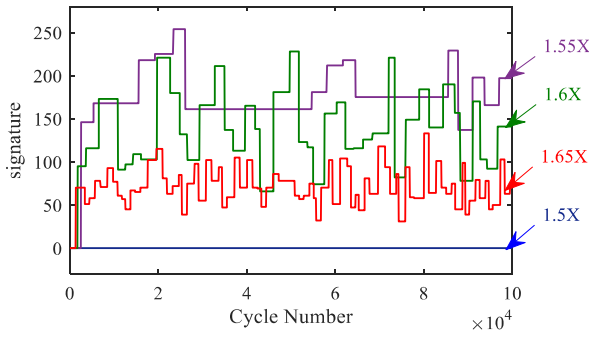

Figure 7 – The signature considering different delay scaling factors.

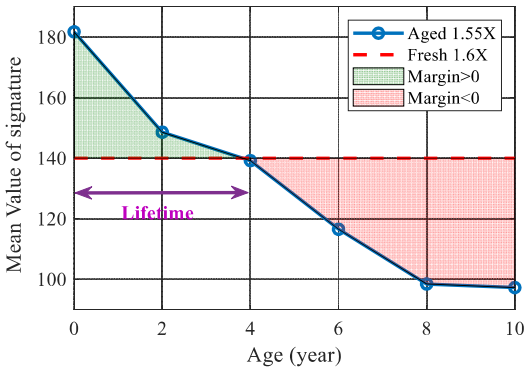

Figure 6 – The averaged signature versus age with *1.55X* delay scaling and the lifetime based on signature with fresh *1.6X* delay scaling.

scaling factor, the overflow frequency of m-bit counter is so low that the timer overflows always and the sampling is not enabled. When the delay scaling factor increases from *1.5X*, the signature generation starts. As can be seen, with a higher delay scaling factor, the signature value is lower and that is because, the overflow frequency of the m-bit counter increases when the delays are scaled up. The waveforms that are shown in Figure 6 verifies that the proposed chip health tracking system reflects the delay degradation with distinct signatures for distinct delay scaling factors.

To get one number which reflects the age of the circuit, the average value of the signature is calculated. In Figure 7, the mean value of the signature over 50K cycles is shown versus the age of the circuit, considering a *1.55X* delay scaling. It can be observed that the age of the circuit is properly reflected in the mean value of the signature. Without loss of generality, consider a speed constrained circuit with scalable supply voltage. Suppose that error free functionality of this circuit is guaranteed when the voltage is scaled down to the point where the delays are*1.6X* more than for the nomnial voltage. The mean value of the corresponding signature (S_L) is shown with the dashed line in the figure. This value is stored when the circuit is fresh and during the lifetime of *the* circuit, the distance between the average signature and the threshold line is calculated as an indicator of available design margins. When the circuit is fresh, the distance is maximum, and as the circuit ages, the disctance decreases, indicating that the slack margin is decreased. When the distance becomes negative, the margins are decreased to a level that the correct functionality of the main circuit cannot be guaranteed anymore. Therefore, the guaranteed lifetime of the circuit is 4 years based on Figure 7.

VI. Conclusions

The proposed technique uses the underlying in-situ delay monitoring hardware more efficiently compared to the conventional static slack monitoring techniques. As the circuit delays degrade, the monitors are excited more frequently. A signature of chip health status is then extracted from the excitation rate of the monitors, which indicates the degradation impact. The proposed idea is implemented for chip health tracking of an ARM Cortex M0 processor. Netlist simulation with aging induced delay degradation verified that the generated signature of chip health status can reflect delay degradation clearly.

References

[1] H. A. Balef et al., "All-Region Statistical Model for Delay Variation Based on Log-Skew-Normal Distribution," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 35, no. 9, pp. 1503-1508, 2016.

[2] A. Vijayan, et al., "Fine-Grained Aging-Induced Delay Prediction Based on the Monitoring of Run-Time Stress," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 37, no. 5, pp. 1064-1075, 2018.

[3] Y. Wang, et al., "Variation tolerant on-chip degradation sensors for dynamic reliability management systems," *Microelectronics Reliability,* vol. 52, no. 9–10, pp. 1787-1791, 2012.

[4] D. Sengupta and S. S. Sapatnekar, "Estimating Circuit Aging Due to BTI and HCI Using Ring-Oscillator-Based Sensors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 36, no. 10, pp. 1688 - 1701, 2017.

[5] L. Lai, et al., "SlackProbe: A Flexible and Efficient In Situ Timing Slack Monitoring Methodology," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD),* vol. 33, no. 8, pp. 1168-1179, 2014.

[6] M. Sadi, et al., "Design of Reliable SoCs With BIST Hardware and Machine Learning," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 25, no. 11, pp. 3237 - 3250, 2017.

[7] M. Nicolaidis, "Time Redundancy Based Soft-Error Tolerance to Rescue Nanometer Technologies," in *Proceedings of the 17th IEEE VLSI Test Symposium (VTS'99)*, Dana Point, 1999.

[8] H. A. Balef, et al., "Effective In-Situ chip health monitoring with selective monitor insertion along timing paths," in *Proceedings of the 28th ACM Great Lakes Symposium on VLSI*, Chicago, 2018.

[9] M. Sadi, L. Winemberg and M. Tehranipoor, "A robust digital sensor IP and sensor insertion flow for in-situ path timing slack monitoring in SoCs," in *Proceedings of the 33rd IEEE VLSI Test Symposium (VTS)*, Napa, 2015.

[10] A. Benhassain, et al., "Investigation of critical path selection for in-situ monitors insertion," in *IEEE 23rd International Symposium on On-Line Testing and Robust System Design (IOLTS)*, Thessaloniki, 2015.

[11] R. Reis, Y. Cao and G. W. (Eds.), Circuit design for reliability, New York: Springer, 2015.