Cross-Layer Design for Reliability in Advanced Technology Nodes: An EDA Perspective

Runsheng Wang*, Zuodong Zhang, Zixuan Sun, Zizheng Guo, Yibo Lin, Ru Huang
School of Integrated Circuits, Peking University, Beijing 100871, China
*Email: r.wang@pku.edu.cn

2022 IEEE 16th International Conference on Solid-State & Integrated Circuit Technology (ICSICT) | 978-1-6654-6906-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSICT55466.2022.9963318

Abstract

Design for reliability (DFR) in advanced technology nodes has become an increasingly challenging task, which requires comprehensive support from aging-aware EDA tools and optimization design flows. In this paper, our recent studies on the cross-layer aging-aware design are summarized, especially from EDA perspectives. The recent advances in reliability modeling device-level, a set of aging-aware modeling and analysis frameworks at the circuit level and system level are overviewed. The results demonstrate that the cross-layer DFR framework enables an accurate analysis to reduce over-design and optimize PPA across lifetime, and can help designers to explore error-resilient architecture more efficiently.

1. Introduction

With CMOS technology scaling to sub-20 nm nodes, the circuit design margin is extremely tightened due to transistor aging and process variation (PV) [1][2][3]. To guarantee the parametric yield and circuit lifetime, designers usually use corner-based analysis to estimate the effects of aging and variation on the circuit, and add a guardband to ensure the reliability of the circuit. However, the overestimation introduced by the corner-based analysis is increasing rapidly with technology scaling, which ultimately obliterates gains from device scaling. Therefore, an efficient and accurate design-for-reliability (DFR) framework is urgently needed.

This paper reviews our recent works in aging-aware design methodology and EDA tools. As shown in Fig. 1, a set of novel cross-layer aging-aware modeling and analysis tools was developed to enable more accurate and efficient aging-aware analysis. Compared with the conventional corner-based analysis, the proposed aging-aware analysis framework can reduce over-design, thus leaving more design margin for designers and optimizing PPA across lifetime. We also demonstrate the potential of cross-layer analysis framework for the design space exploration of error-resilient system.

2. Device-level Modeling

It is well known that transistor aging has become more pronounced due to the slow scaling of power supply voltages, which, combined with shrinking device dimensions, leads to increasing vertical and horizontal electric fields. These high fields lead to phenomena like hot carrier degradation (HCD) and negative bias temperature instability (NBTI) both of which cause transistor degradation.

Recent works about NBTI mainly focus on the full range of degradation and recovery effects, from short-term to long-term, as well as their voltage, temperature [4][5], frequency and body bias dependences [6]. Furthermore, the history effect of NBTI is non-negligible,

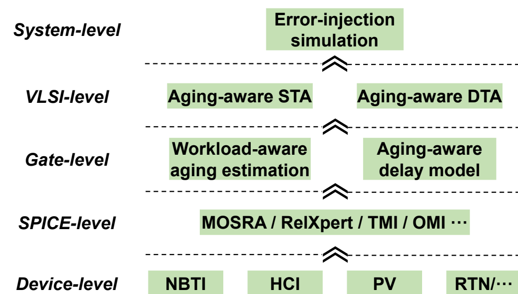

Figure 1. The overview of the cross-layer reliability-aware analysis framework. The framework includes commercial tools, open source projects, and our in-house tools and models

as it can cause abnormal over-/under-shoot when the circuit operated under dynamic voltage scaling [4].

Recent studies on HCD are the proposed multi trap-based compact model [7], which can accurately predict hot carrier degradation and variation in the full $\{V_{gs}, V_{ds}\}$ bias region. The microscopic speculation of interface generation was revealed [8]. Based on the above, the typical locations of HCD generated traps in planar MOSFETs [9] and FinFETs [10] were demonstrated, respectively. Also, the non-universal temperature dependence of HCD was investigated [11]. In addition, the mixed-mode reliability of HCD-BTI coupling through self-heating and under off-state stress is investigated [8].

3. Circuit-level Analysis

The stress mode of transistors is different in digital circuits and analog circuits. Therefore, the main degradation effect is also different. In the digital circuit, the toggle rate is usually low and the transition time is small, hence the static vertical electric field stress is the main aging stress, which means NBTI dominates the transistor aging. While the stress modes of transistors in an analog circuit are more diverse, therefore, HCD is also non-negligible.

3.1 Aging-aware Analysis for Analog Design

Fig. 2 shows the SPICE-level reliability simulation flow [4]. The inputs of the flow are the circuit, the testbench, and the test stimulus. In addition, a device-level aging model is required. The analysis is performed in two stages: stress and aged. First, a fresh SPICE transient simulation is performed and the stress waveforms are recorded. Then, the device-level reliability model is adopted to calculate the degraded parameters. Finally, a post-stress simulation is performed again to obtain the aged circuit results. Because all waveforms are recorded, this flow is compatible with most of the device-level degradation models.

Other effects that require special attention include

Authorized licensed use limited to: Peking University. Downloaded on March 06,2026 at 06:27:31 UTC from IEEE Xplore. Restrictions apply.

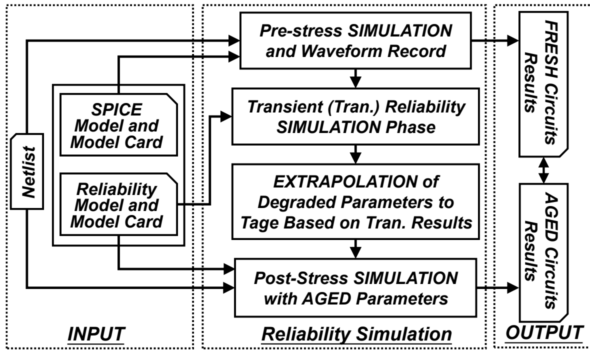

Figure 2. The circuit reliability simulation flow with SPICE reliability interface.

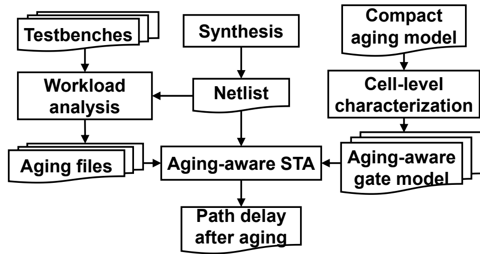

Figure 3. A general overview of the aging-aware STA flow.

accelerated aging in analog circuits with feedback. For example, AMS circuits are typically biased in the saturation region during power ON mode. However, such conditions lead to continuous HCD degradation leading to a shift in threshold voltage as well as circuit parameters over time. Dynamically varying operating voltages influences the aging rate of devices and under such feedback configurations, the degradation rate can be significantly enhanced and cause sudden failure of the circuit. Therefore, an iteration-based lifetime estimation to capture the aging effect in circuits with feedback has been proposed [12].

3.2 Aging-aware Analysis for Digital Design

The scale of digital circuits is usually much larger than that of analog circuits. Therefore, SPICE simulation is not practical for digital circuit analysis. The conventional digital circuit design flow builds gate-level libraries to support large-scale gate-level timing analysis. To enable the aging-aware design for large-scale digital circuits, the aging-aware gate-level model and timing analysis tool are key parts.

Fig. 3 illustrates a general flow of the aging-aware static timing analysis (STA), including the workload analysis and the aging-aware delay model. The workload analysis estimates the aging of each transistor under specific working conditions (including but not limited to temperature, voltage, frequency, and tasks performed). The aging-aware delay model calculates the aged delay and transition time of each gate after aging.

3.2.1 Workload Analysis

The workload analysis is implemented in two steps [13]: In the first step, we use the zero-delay gate-level simulation to obtain the static probability and toggle rate of each internal net; then, we employ the cell-level analytical model and device-level aging model to calculate the ΔVth of each transistor. For example, the

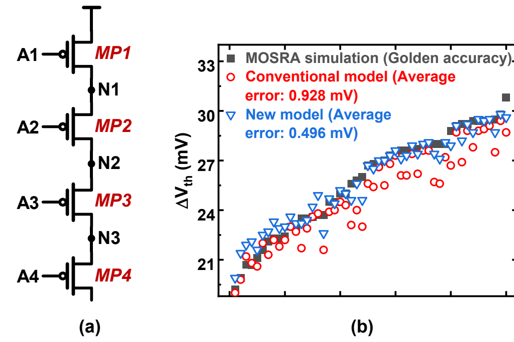

Figure 4. (a) Schematic of *NOR4* gate and (b) NBTI degradation of MP3.

stress probability (SP) of the PMOS in the NOR4 gate can be determined as the following equations (the transistor-level schematic of the NOR4 gate is shown in Fig. 4 (a)):

$$SP(MP1) = P(A1 = 0)$$

$$SP(MP2) = P(A2 = 0) \cdot [P(A1 = 0) + P(A1 = 1) \cdot P(A3\ or\ A4 = 1) \cdot P(A1 = 0)]$$

$$SP(MP3) = P(A3 = 0) \cdot [P(A1 = 0) \cdot P(A2 = 0) + P(A4 = 1) \cdot P(A1\ or\ A2 = 1) \cdot P(A1\ and\ A2 = 0)]$$

$$SP(MP4) = P(A1 = 0) \cdot P(A2 = 0) \cdot P(A3 = 0) \cdot P(A4 = 0)$$

The first item of the equations is the probability of $V_d = V_{dd}$ and $V_g = 0$, which represents the NBTI stress from V_{dd}. The second item is the probability of the source and drain in the floating state and the voltage is high, which represents the NBTI stress from the floating state. As shown in Fig.4 (b), the new model including floating state stress achieves higher accuracy than the conventional model which only considers the stress from V_{dd}.

3.2.2 Aging-aware Delay Model

The other essential part is the aging-aware delay model. Compared with the traditional delay model, the aging-aware delay model requires an extra input: the degradations of each transistor in the cell, inevitably making the model more complex. An industry-friendly aging-aware delay model should simultaneously meet the requirements of accuracy, characterization complexity, and model size at the same time. Therefore, in our framework, we use the first-order Taylor expansion to model the aging effect on the delay and transition time (tr):

$$delay_{aged} = delay_{fresh} + \sum_{i \in I} a_i \Delta V_{thi}$$

$$tr_{aged} = tr_{fresh} + \sum_{i \in I} b_i \Delta V_{thi}$$

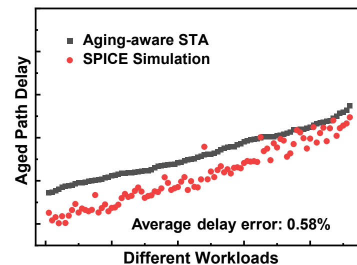

Figure 5. Path delay calculated by aging-aware STA and SPICE simulation under different workloads.

Authorized licensed use limited to: Peking University. Downloaded on March 06,2026 at 06:27:31 UTC from IEEE Xplore. Restrictions apply.

where I is the set of all transistors in the cell, $delay_{fresh}$ and tr_{fresh} are obtained from the standard cell libraries. And ΔVth is the threshold voltage shift due to aging.

According to the SPICE simulation, the sensitivity coefficients *a, b*, vary with different input slews and output loads, so we use the Ridge Regression model to fit the dependency:

$$a_i = a_{i0} + \alpha_{ai} \cdot slew + \beta_{ai} \cdot load$$

And other sensitivity coefficients are fitted in the same way. For a cell containing N transistors, the proposed model only needs to store 12N aging parameters, which reduces the model size significantly. To further reduce the characterization complexity, two other approaches are also proposed [14].

4. EDA Tools Developments

4.1 Aging-aware STA

Based on the gate-level models discussed in section 3.2, the workload-dependent aging-aware STA flow can be implemented. To evaluate the accuracy of the proposed aging-aware gate-level model, multiple ISCAS'85 benchmark circuits are used to perform the aging-aware path-based STA. And the reported paths are also recalculated by the SPICE simulation as the golden accuracy.

Fig. 5 shows the aged delay of a critical path in C432 circuit under 100 different workloads. According to the SPICE simulation results, there is a notable variation of aging-induced delay between different workloads. It can be seen that the workload dependency is well captured by the proposed model.

Table I lists the critical path delay degradation of various ISCAS' 85 benchmark circuits. The aged delays are average values for different workloads. It can be seen that, for all benchmark circuits, the path delay calculated by the proposed aging-aware model is very close to that obtained from the SPICE simulation, regardless of the scale of the circuit. On average, the proposed aging-aware delay model introduces only 0.5% pessimism, which means that the proposed model can reduce the pessimism, and thus largely save the precious design margin. The results prove that the proposed aging-aware model is accurate enough for most design scenarios.

4.2 Aging-aware DTA

Recently, to improve energy efficiency, many optimization techniques usually go beyond the static timing constraints, and leverage dynamic timing information, such as application-based dynamic-voltage-frequency-scaling (DVFS) [15]. It is based on the observation that an instruction executed on an embedded processor may not utilize all the functionalities. Therefore, a timing slack may exist between the most critical path reported from STA and the longest path triggered by the instruction. Distinguish from previous static timing slack (STS), such timing slack is called dynamic timing slack (DTS). The presence of DTS means that for the specific application, the V_{dd} or clock cycle could have been smaller. Therefore, using application-specific V_{min}/f_{max} instead of the fixed V_{dd}/f_{max} can increase energy efficiency.

The calculation of DTS relies on dynamic timing analysis (DTA). However, existing DTA tools adopt corner-based analysis to cover the impact of aging and variation, and the pessimism of the worst-case corner ultimately leads to an over-designed system.

Therefore, we proposed AVATAR, an aging-&variation-aware dynamic timing analyzer [16], based on an event-based DTA algorithm [18] and the aging-aware gate-level model discussed in the previous section. The overall task flow of AVATAR is shown in Fig. 6. It contains two parts: the gate-level aging and variation

Table 1. Path Delay Degradation of Benchmark Circuits

Design	Fresh Delay (ps)	Aged Delay (ps)		Pessimism
		SPICE simulation	Proposed model	
C432	373.98	401.27	403.45	0.58%
C880	371.45	402.73	404.97	0.61%
C1355	312.39	331.86	333.04	0.38%
C1908	457.17	488.81	490.47	0.36%
C2670	446.14	479.00	482.32	0.74%
C3540	529.96	564.87	566.56	0.32%
C5315	465.61	495.68	497.07	0.30%
C6288	1155.6	1243.04	1246.95	0.34%
C7552	838.83	897.22	903.98	0.81%

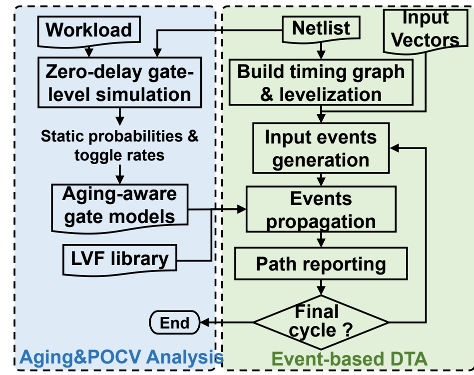

Figure 6. The proposed aging- and variation-aware DTA flow.

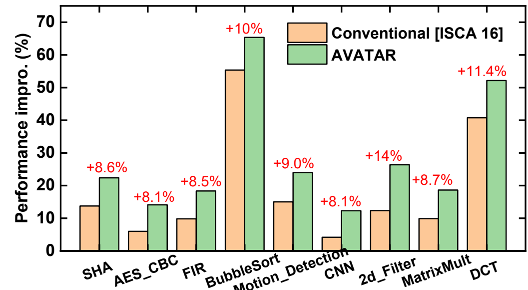

Figure 7. Performance improvement from application-based DVAFS based on the corner-based DTA and AVATAR.

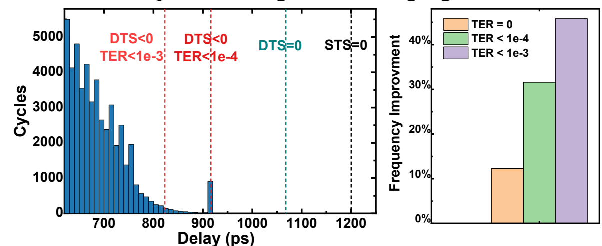

Figure 8. (a) Histogram of dynamic delays per cycle for the RISC-V core running CNN application. (b) The performance gains by allowing a small number of timing errors.

Authorized licensed use limited to: Peking University. Downloaded on March 06,2026 at 06:27:31 UTC from IEEE Xplore. Restrictions apply.

analysis and the event-based DTA.

The gate-level aging analysis is the same as that in aging-aware STA. For random variation analysis, we adopt parametric on-chip variation (POCV) analysis [17]. POCV can be represented with the liberty variation format (LVF). To reduce the characterization complexity, we model aging and random variation as two independent effects and only consider the effect of aging on the means value of cell delay and transition time.

We target application-based DVFS as one of the use cases of AVATAR. An open RISC-V core and a set of embedded application benchmarks [19] are used to validate the DVFS flow. We use two methods to determine the application-specific V_{min}/f_{max}: 1) Using the corner-based DTA with extra aging guardband and variation guardband in dynamic delay calculation. 2) Using AVATAR. The final delay is calculated as $\mu(delay) + 3 \times \sigma(delay)$.

The performance improvement by adopting the performance-first DVFS strategy is shown in Fig. 7. The results show that different applications can expose different maximum dynamic delays, depending on which paths are triggered. The flow based on the corner-based DTA can improve the performance by an average of 17.13%, while the flow based on AVATAR can improve the performance by an average of 26.59%. The additional improvement of AVATAR over the corner-based DTA varies from 8% to 14%.

Another advantage of DTA is its capability to estimate the timing error rate, which allows designers to tradeoff computing accuracy with performance or efficiency. Fig. 8 (a) shows the histogram of the dynamic delay for RISC-V core running *CNN* application. It can be seen that although *CNN* can trigger some long paths, the trigger rate is rather small. This means that a significant amount of power and performance needs to be sacrificed to prevent a small probability of error. On the other hand, the previous work shows that *LeNet-5* can tolerate error rates of 1e-4 to 1e-3 without affecting the network accuracy [20]. Therefore, the minimal clock cycle can be much smaller if we relax the constraint of computing accuracy. As shown in Fig. 8 (b), allowing timing errors to occur can lead to up 45.8% performance improvement.

5. Summary

In this paper, we review our recent works on the cross-layer aging-aware analysis framework. New gate-level aging-aware models are explored, which enable more accurate aging-aware timing analysis for large-scale digital circuits. The optimization strategy based on new dynamic timing analysis tools is evaluated. The results indicate that cross-layer design is a promising solution for further improving efficiency at advanced technology nodes.

Acknowledgments

This work was supported in part by the National Key R&D Program (2020YFB2205500), NSFC (62125401) and the 111 Project (B18001). R. Wang thanks J. Wang, D. Wu, J. Xie and T. Guo for the collaborative projects, and former students S. Guo, Z. Zhang, Z. Yu and J. Zhang for the input.

References

[1] R. Huang et al., "Variability-and reliability-aware design for 16/14nm and beyond technology," International Electron Devices Meeting, 2017, pp. 12.4.1-12.4.4.
[2] R. Wang et al., "Too Noisy at the Bottom? —Random Telegraph Noise (RTN) in Advanced Logic Devices and Circuits," International Electron Devices Meeting, 2018, pp. 17.2.1-17.2.4.
[3] R. Wang et al., "Can Emerging Computing Paradigms Help Enhancing Reliability Towards the End of Technology Roadmap?," International Reliability Physics Symposium, 2021, pp. 1-7.
[4] S. Guo et al., "Towards reliability-aware circuit design in nanoscale FinFET technology: New-generation aging model and circuit reliability simulator," International Conference on Computer-Aided Design, 2017, pp. 780-785.
[5] Z. Zhang et al., "Circuit Reliability Comparison Between Stochastic Computing and Binary Computing," Transactions on Circuits and Systems II, pp. 3342–3346, 2020.
[6] J. Zhang et al., "Body Bias Dependence of Bias Temperature Instability (BTI) in bulk FinFET Technology" Energy & Environmental Materials, pp. 1-4, 2022.
[7] Z. Yu et al., "Hot Carrier Degradation-Induced Dynamic Variability in FinFETs: Experiments and Modeling," Transactions on Electron Devices, pp. 1517-1522, 2020.
[8] R. Wang et al., "Understanding Hot Carrier Reliability in FinFET Technology from Trap-based Approach," International Electron Devices Meeting, 2021, pp. 31.2.1-31.2.4.
[9] Z. Sun et al., "Investigation on the Lateral Trap Distributions in Nanoscale MOSFETs During Hot Carrier Stress," IEEE Electron Device Letters, pp. 490-493, 2019.
[10] Z. Yu et al., "On the Trap Locations in Bulk FinFETs After Hot Carrier Degradation (HCD)," in IEEE Transactions on Electron Devices, pp. 3005-3009, 2020.
[11] Z. Yu et al., "Non-Universal Temperature Dependence of Hot Carrier Degradation (HCD) in FinFET: New Observations and Physical Understandings," Electron Devices Technology and Manufacturing Conference , 2018, pp. 34-36.
[12] K. B. Sutaria et al., "Accelerated Aging in Analog and Digital Circuits With Feedback," Transactions on Device and Materials Reliability, pp. 384-393, 2015.
[13] Z. Zhang et al., "Aging-Aware Gate-Level Modeling for Circuit Reliability Analysis," Transactions on Electron Devices, pp. 4201–4207, 2021.
[14] X. Zhang et al., "Efficient Aging-Aware Standard Cell Library Characterization Based On Sensitivity Analysis," Transactions on Circuits and Systems II, submitted.
[15] H. Cherupalli et al., "Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems," International Symposium on Computer Architecture, p. 671–681, 2016.
[16] Z. Zhang et al., " AVATAR: An Aging- and Variation-Aware Dynamic Timing Analyzer for Application-based DVAFS," Design Automation Conference, 2022.
[17] A. B. Kahng, "New game, new goal posts: A recent history of timing closure," Design Automation Conference, 2015, pp. 1–6.
[18] Z. Zhang et al., "EventTimer: Fast and Accurate Event-Based Dynamic Timing Analysis," Design, Automation & Test in Europe Conference & Exhibition, 2022, pp. 945-950.
[19] M. Gautschi et al., "Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices," Transactions on Very Large Scale Integration Systems, pp. 2700-2713, 2017.
[20] X. Jiao et al., "An assessment of vulnerability of hardware neural networks to dynamic voltage and temperature variations," International Conference on Computer-Aided Design, 2017, pp. 945-950.

Authorized licensed use limited to: Peking University. Downloaded on March 06,2026 at 06:27:31 UTC from IEEE Xplore. Restrictions apply.