Reliability-Enhanced Circuit Design Flow Based on Approximate Logic Synthesis*

Zuodong Zhang, Runsheng Wang, Zhe Zhang, Ru Huang Institute of Microelectronics, Peking University, Beijing, China r.wang@pku.edu.cn Chang Meng, Weikang Qian University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, China qianwk@sjtu.edu.cn Zhuangzhuang Zhou Electrical and Computer Engineering Department, Cornell University, Ithaca, NY 14853, USA

ABSTRACT

With the downscaling of CMOS technology, the circuit design margin becomes more and more tight due to wider guardband, which is required to counteract the severer transistor aging and variations. Thus, reliability-enhanced circuit design is urgently needed to reduce the guardband. In this paper, a reliabilityenhanced design framework based on approximate synthesis is proposed to completely eliminate the aging guardband. It mainly includes two key parts: first, a forward reliability simulation flow supporting statistical static timing analysis (SSTA) is performed to estimate the path failure rates after aging; if the timing constraints are not satisfied, then a backward delay-driven approximate logic synthesis flow will perform approximate local changes on the critical paths to reduce the delay until the reliability requirement is finally satisfied and no aging guardband is needed. The results show that the approximate circuit has a smaller aged delay than the original circuit, so that the path failure rates are significantly decreased. It indicates that the proposed design flow can convert the timing errors that have fatal impact on applications, into negligible error on low-significance bits to improve the resilience of circuits, which provides a new perspective of reliability-enhanced design at nanoscale.

KEYWORDS

Reliability-enhanced design; Approximate computing; Circuit reliability simulation; Negative bias temperature instability (NBTI); Guardband; Aging; Statistical static timing analysis (SSTA); Logic synthesis

GLSVLSI '20, September 7–9, 2020, Virtual Event, China © 2020 Association for Computing Machinery. ACM ISBN 978-1-4503-7944-1/20/09\$15.00 https://doi.org/10.1145/3386263.3406926

ACM Reference format:

Zuodong Zhang, Runsheng Wang, Zhe Zhang, Ru Huang, Chang Meng, Weikang Qian, and Zhuangzhuang Zhou. 2020. Reliability-Enhanced Circuit Design Flow Based on Approximate Logic Synthesis. In *GLSVLSI* '20, September 7–9, 2020, Virtual Event, China. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3386263.3406926

1 INTRODUCTION

With CMOS technology continuously shrinking, the reliability issues have become more and more severe [1,2]. The larger process variations and transistor aging effects such as negative bias temperature instability (NBTI) make it difficult to guarantee the circuit lifetime in advanced technology nodes. To ensure the circuit performance at the end of life, larger frequency/voltage guardband is needed, which will reduce the circuit speed and/or increase the power. Thus, the benefits of technology scaling are sacrificed.

In order to relieve the impact of reliability issues, solutions from device and circuit perspectives have been proposed, such as aging-aware voltage and frequency scaling [3,4] and aging sensor [5]. In essence, all these methods aim to accurately evaluate the required aging guardband, instead of reducing or removing it. Other solutions like aging control gate [6] or resizing [7], can reduce the guardband by altering the circuit structure, but they will bring additional overhead in area and power. On the other hand, whether it is possible to take advantage of the new computing paradigm to enhance the circuit reliability or even completely remove the aging guardband, also deserves study. In Ref. [8], a circuit reliability comparison between stochastic computing and binary computing shows that the performance of stochastic computing circuits is intrinsically resistant to variabilities due to the circuit topology and the probability encoding.

Approximate computing is a promising computing paradigm that has attracted more and more attention in recent years [9]. The purpose of approximation is to intentionally introduce some errors that have small effects at the application level, in exchange for higher speed and/or smaller area. It has been demonstrated that approximate computing can improve energy efficiency and area efficiency in many systems and applications that can tolerate some loss of quality or optimality in the result, such as deep neural

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

networks, data mining, and video/image processing [10]. However, most researches on approximate computing in the past focused on improving the area efficiency without paying attention to the reliability of the approximate circuit. In Ref. [11], a method to enhance reliability by truncating the bits of adders was proposed. However, it did not propose a general design method, and the truncation is not an optimal approximation [12].

In this paper, a reliability-enhanced circuit design method based on approximate logic synthesis (ALS) is proposed. The main contributions of this work are as follows:

(1) Two reliability simulation methods supporting statistical static timing analysis (SSTA) are proposed, based on two different workload analysis methods. We evaluate the accuracy and the efficiency of both methods and propose a design flow that combines the advantages of both.

(2) A delay-driven ALS algorithm is adopted, which can reduce the delay of critical paths by deploying approximate local changes (ALCs).

(3) A reliability-enhanced circuit design framework is proposed, which contains the above forward reliability simulation flow and the backward ALS algorithm. It can trade off accuracy with higher speed and longer lifetime. The result shows that the proposed reliability-enhanced approximate (REA) circuit can work well without aging guardband after 10 years of aging.

2 PRELIMINARY

2.1 NBTI and Aging Guardband



Figure 1: The degradation and recovery effect of NBTI under AC workload with duty factor (DF) α .

Process variations and transistor aging are two of the most important factors affecting circuit reliability. In digital circuits, the NBTI dominates the transistor aging [3], which originates from trapping accumulation in PMOS during the circuit operation.

The speed of trapping accumulation depends on stress voltage and temperature. Because of the recovery effect, as shown in Fig. 1, the NBTI in a circuit also depends on the working frequency and duty factor (DF). In digital circuit, the waveform of each node can be considered as a square wave, so the aging of each MOSFET mainly depends on the working frequency and the corresponding DF. In digital circuit, the shift of V_{th} causes a reduction of the drain current of PMOS, resulting in slower charging of capacitive load, which means that the delay of output rising edge increases in gate level. Although digital circuit is more resilient, degradation may still cause the increase of path delay and the circuit may violate the setup check. Therefore, in order to ensure the lifetime of the circuit, an aging guardband is usually added in timing analysis, and the guardband is over 15% in advanced technology nodes [22].

2.2 Approximate Computing

Approximate computing is an emerging circuit design paradigm, which modifies the function of the target circuit while ensuring the usability of the application. The approximate circuit has smaller area, lower power, and higher speed than the original circuit. As shown in Fig. 2, comparing with the traditional mirror full adder, the approximate full adder [13] reduces the number of MOSFET from 24 to 11, thus the overall area is 0.36 of the original circuit and the power consumption is 0.175 of the original one. In addition to the approximate full adder, researchers also consider approximate carry lookahead adder [14], approximate multiplier [15], etc.



Figure 2: (a) Conventional full adder and (b) approximate full adder.

As a new design paradigm, the key technique driving approximate computing to practical application is ALS [16]. ALS aims to find the optimal approximate circuit satisfying an error constraint, such as the error rate (ER) constraint or the mean error distance (MED) constraint. In most of the previous works, the optimization goal is to minimize area. A recent work [17] proposes a delay-driven approximate logic synthesis (DALS) framework, in which the optimization goal is the circuit delay.

3 RELIBILITY-ENHANCED DESIGN FLOW

Given that logic approximation can reduce the delay, a reliability-enhanced approximate design flow is proposed, as shown in Fig. 3. The proposed flow consists of two key procedures: forward reliability simulation and backward ALS. The forward reliability simulation supports SSTA, so that the path failure rates (PFRs) can be estimated. If the largest PFR is higher than the given threshold, it means that the lifetime of the target circuit does not meet the design requirements. In this case, the circuit needs to be processed backward by approximate synthesis to cut the logic depth by one layer. After that, the reliability is checked again. This procedure is repeated iteratively until the reliability requirements are satisfied. In order to accelerate the iteration, we propose two methods of reliability simulation: one is to estimate the aged delay quickly in iteration, the other is to re-evaluate the PFR accurately after iteration. In this way, it can not only accelerate the iteration speed by several times, but also guarantee the lifetime of the approximate circuit. The specific procedure of the two reliability simulations will be introduced in the next section.



Figure 3: Flow diagram of the proposed reliabilityenhanced design flow based on approximate synthesis.

The approximate algorithm works on the AND-inverter graph (AIG) [23]. In each iteration, the algorithm finds the best approximate local changes to reduce the depth of the critical path. Note that the depth reduction in an AIG usually results in a delay reduction in the final mapped circuit. To ensure the usability of the approximate circuit in the application, an error constraint is also needed. If the error of approximate circuit reaches the limit while the PFR is still above the threshold, a timing guardband must be added. In essence, our method sacrifices the computing accuracy at low-significance bits in exchange for circuit reliability improvement. In other words, the timing violations that seriously affect the circuit function are converted into deliberately induced errors that have negligible impact in practical applications.

4 CIRCUIT RELIABILITY EVALUATION

The goal of reliability simulation is to analyze the timing after circuit aging. It is mentioned that NBTI in digital circuits depends on the working frequency and the DF of each transistor. Therefore, the reliability simulation is divided into three parts: workload analysis, transistor aging calculation, and timing analysis after aging.

In this work, two reliability simulation methods are proposed, which are used for fast estimation and accurate re-evaluation. The corresponding frameworks are shown in Fig. 4 and Fig. 5, respectively. The main difference between the two methods is the workload analysis part, which is the most time-consuming part.

4.1 Workload Analysis

The workload analysis is to obtain the DF of each transistor by simulating the actual operating condition as accurate as possible. The most accurate method is SPICE-level simulation of the whole netlist with input testbench. However, since the simulation complexity increases as the circuit scale increases, it is impossible to simulate VLSI circuit at SPICE level. Therefore, two simplified methods are proposed below.



Figure 4: Flow diagram of the proposed fast reliability simulation flow based on analytical DF calculation.



Figure 5: Flow diagram of the proposed accurate reliability simulation flow based on SPICE-based DF calculation.

The first is analytical DF calculation, which only calculates the DF of each node without record the actual waveform. The basic principles are shown in Fig. 6. For each node in the circuit, DF can be calculated based on its connected logic gate. For example, the output DF of an AND gate is the product of two inputs' DFs. After setting the DFs of input signals, the DFs of the internal nodes in a circuit can be calculated in a topological order. With those DF values known, we can further get the DF for each MOSFET in a gate. It is defined as the ratio of stress time to total runtime, which can be calculated by the DF of each input and the stack structure.

For example, the PMOS A is stressed when the input A is 0, while PMOS B is stressed when the inputs A and B are both 0.

The benefit of analytical calculation is obvious. Because complex circuit simulation is replaced by simple probability propagation, the time consumed is reduced by several orders. However, this method has two disadvantages. One is that it ignores the correlation of the signals caused by the reconvergent path of the circuit, and the other is the stack effect. The stack effect occurs in a series structure, like the PMOS in a NOR4 gate. As shown in Fig. 7, when the PMOS A and D are turned off, the nodes in the middle are floating, so PMOS B and C may also be in stress. A comparison between SPICE simulation and analytical DF calculation is also shown in the figure, where the DF of all input signals are assumed to be 0.5. The result shows that the DF of B and C are higher than that calculated by the formula, because even if A is turned off, the voltage of middle nodes may still be high. Stack effect can lead to underestimation of aging in DF calculation method, so we also propose a path-based SPICE level workload analysis.



Figure 6: Calculation formula of DF for (a) each circuit nodes and (b) each MOSFET.



Figure 7: Stack effect will affect the accuracy of DF calculation.

The second method we proposed is to use both gate-level and SPICE-level simulation to calculate the DF of each transistor in critical paths. The SPICE simulation of the entire circuit netlist is too expensive, but it is very easy to simulate only several paths. Therefore, as shown in Fig. 5, the workload analysis can be divided into two steps. First, the N-worst paths are found by fresh static timing analysis, and the waveform of each external nodes on these paths are obtained by gate-level simulation. Then, the DF of each MOSFET is calculated by SPICE simulation with open model interface (OMI) or TSMC model interface (TMI). In this way, the accurate workload can be analyzed to support the reliability evaluation of VLSI circuit

4.2 Long-Term Transistor Aging Model

In Ref. [18], an accurate NBTI degradation and recovery model was proposed, which is suitable for arbitrary waveform. In largescale digital circuit, the model can be simplified by assuming that the waveforms of each node are approximately square waves, shown by the following equations:

$$\Delta V_{th_fast} = M \cdot A_1 \exp(B_1 V_g) \exp\left(-\frac{E_{a_1}}{k_B T/q}\right) \log(1 + C \cdot T \cdot DF)$$
(1)

$$\Delta V_{th_slow} = A_2 \exp\left(B_2 V_g\right) \exp\left(-\frac{E_{a2}}{k_B T/q}\right) (t \cdot DF)^{n_1} \quad (2)$$

$$\Delta V_{th} = \Delta V_{th_slow} + \Delta V_{th_fast} \tag{3}$$

$$\sigma^2(\Delta V_{th}) = \eta \cdot \mu(\Delta V_{th}) \tag{4}$$

where M is the modulation factor, T is the period, t is the aging time and DF is the equivalent DF. The new long-term model needs the DF of each transistor and the working frequency of the whole circuit, which is suitable with the above workload analysis method.

As shown in Fig. 8, this long-term aging model agrees well with 16/14 nm FinFET experimental data, in which the frequency dependency and DF dependency are well captured.



Figure 8: The comparison of experimental results with the proposed long-term model under AC stress with different (a) frequencies and (b) duty factor.

4.3 Statistical Static Timing Analysis

As transistor shrink into nanoscale, the impact of process variations and aging variations cannot be ignored. If designers continue to use the worst-case corner method, the design margin will be extremely tight, because the worst-case design always overestimates the total impact of variations. Therefore, statistical analysis is urgently needed to relax the design margin. In the traditional timing analysis framework, a statistical aging library is needed to implement the statistical timing analysis after degradation. However, such library is very difficult to build.

Therefore, a new SSTA framework is proposed, as shown in Figs. 4 and 5. The basic idea is that, rather than build a complex library, we perform a fresh STA at first. After obtaining the slope

and load of each gate on critical paths, we use Monte Carlo (MC) simulation to analyze the degraded delay distribution. Although this method will increase the simulation time for timing analysis, considering that the length of critical paths is usually not too long, the MC simulation time is acceptable.



Figure 9: Illustration of delay-driven approximate logic synthesis algorithm.

5 APPROXIMATE SYNTHESIS

Approximate synthesis is a general method for approximate circuit design. In the proposed REA design framework, the approximate synthesis finds the optimal approximate local changes, which can reduce the delay and have the least error impact. Our synthesis algorithm works on the AIG representation of a circuit, and the basic procedure of the algorithm is shown in Fig. 9. In AIG representation, the delay of a circuit is proportional to the depth of AIG. The subgraph containing all critical paths is defined as the critical graph. To reduce the delay of the circuit, the depth of the critical graph needs to be reduced. Because there are usually more than one critical path, multiple approximate local changes need to be performed simultaneously on a cut of the critical paths, which is defined as a critical cut. For example, the cut with nodes 8 and 9 is a critical cut. Therefore, the problem of the approximate algorithm can be described as how to find the optimal critical cut in the critical graph, which has the minimum error impact on the circuit.

The most straightforward approach is to enumerate all sets of ALC and critical cut and then choose the combination with the minimum error. However, the total number of combinations increases exponentially with the size of the graph. To reduce the complexity, it proposes to transform this problem into a network flow problem. First, it enumerates the error impact of applying a set of ALCs. Then, it maps the original critical graph into a critical error network (CEN). Finally, it obtains the minimum cut of the CEN by solving a maximum flow problem on CEN. The minimum cut gives the optimal critical cut to be approximated.

6 EXPERIMENTAL RESULTS

In this section, the experimental results of the proposed REA design method are presented. The benchmark applications are discrete cosine transformation (DCT) and inverse discrete cosine transformation (IDCT), which are usually deployed in multimedia

design to encode and decode images or videos. We deploy an 8bit multiplier and a 16-bit adder. We used Yosys open synthesis suit [19] to synthesize the Verilog HDL codes to BLIF files as the inputs to our in-house program. Technology mapping is performed by the logic synthesis tool ABC [20] using the MCNC generic standard cell library [21]. MED is used as the error metric, which is a widely-used error metric for approximate arithmetic circuits.



Figure 10: PSNR of circuits designed by different approximate method at the different approximate levels.



Figure 11: The fresh and aged path delay of the circuit under different approximate level and the required guardband.

Fig. 10 shows the peak signal to noise ratio (PSNR) at the different approximate levels of the 16-bit adder (approximate level is the reduced logic depth). The results indicate that, our general synthesis method can provide a finer level of approximation than simply truncating the low-significance bits [11], which is more friendly to the low-precision arithmetic circuit. The better circuit approximation is because our synthesis algorithm always adopts the ALC that has the least error impact, while truncating is not the best ALC in most cases.

Fig. 11 shows the mean circuit delay before and after aging at the different approximate levels of the adder, as well as the required guardband under these levels. The mean and variance of the delay of critical path are obtained from accurate reliability simulation, and the guardband is calculated by $3-\sigma$ standard. As mentioned before, a cut on AIG does not necessarily means a significant reduction in critical path delay, but the delay is monotonically decreasing with the approximate level. When the approximate level is greater than 3, in other words, after 3 iterations of the design flow, the timing guardband is no longer needed, which means that the timing guardband can be completely eliminated using logic approximation.



Figure 12: Images processed by the fresh original circuit (a), fresh REA circuit (b), aged original circuit (c) and aged REA circuit (d)

Based on the results of SSTA and the timing constraints, the failure rate of each path can be calculated. Fig. 12 shows the output picture of the DCT-IDCT circuit. After 10 years of aging, the image quality of the original circuit is greatly reduced due to timing error. Because timing errors are more likely to occur on longer paths, that is, the more significant bits of the adder, it will seriously affect the computing result. In contrast, for the circuit with REA design, although its initial PSNR decreases slightly (less than 1 dB), its performance after degradation does not decrease due to the shortening of the critical path. On the other hand, the logic approximation starts from the low-significance bit first, so the drop in PSNR is relatively small. The result proves that the proposed REA design can transform timing errors to improve resilience.

7 CONCLUSION

In this paper, a reliability-enhanced circuit design framework is proposed, which contains forward reliability simulation flow and backward approximate logic synthesis. We test our design flow in DCT-IDCT circuit design. The result shows that the aged path failure rate of the approximate circuit is much lower than the original circuit, which demonstrates that the proposed approximate logic synthesis framework can enhance the robustness of circuit and completely eliminate the aging guardband. This work provides a new perspective in reliable circuit design, especially for error-tolerant applications: the computing precision can be traded off with higher speed and longer lifetime. It also suggests that the cross-layer design framework is particularly needed in advanced technology nodes.

ACKNOWLEDGMENTS

This work is partly supported by the National Key R&D Program of China (2020YFB2205502), NSFC (61874005, 61421005, 61927901) and the 111 Project (B18001).

REFERENCES

- R. Huang et al., "Variability-and reliability-aware design for 16/14nm and beyond technology," in 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2017, pp. 12.4.1-12.4.4.
- [2] Z. Ji, H. Chen *et al.*, "Design for reliability with the advanced integrated circuit (IC) technology: challenges and opportunities," *Sci. China Inf. Sci.*, vol. 62, no. 12, p. 226401, Dec. 2019.
- [3] Y. Cao et al., "Cross-Layer Modeling and Simulation of Circuit Reliability," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 33, no. 1, pp. 8-23, Jan. 2014.
- [4] C. Chen et al., "System-Level Modeling of Microprocessor Reliability Degradation Due to Bias Temperature Instability and Hot Carrier Injection," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 8, pp. 2712-2725, Aug. 2016.
- [5] J. P. Kulkarni et al., "A 409 GOPS/W Adaptive and Resilient Domino Register File in 22 nm Tri-Gate CMOS Featuring In-Situ Timing Margin and Error Detection for Tolerance to Within-Die Variation, Voltage Droop, Temperature and Aging," in *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 117-129, Jan. 2016.
- [6] S. Arasu et al., "Controlling Aging in Timing-Critical Paths," in IEEE Design & Test, vol. 33, no. 4, pp. 82-91, Aug. 2016.
- [7] M. Ebrahimi et al., "Aging-aware logic synthesis," in 2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Jose, CA, 2013, pp. 61–68.
- [8] Z. Zhang et al., "Circuit Reliability Comparison between Stochastic Computing and Binary Computing," in IEEE Transactions on Circuits and Systems II: Express Briefs, doi: 10.1109/TCSII.2020.2993273.
- [9] J. Han et al., "Approximate computing: An emerging paradigm for energyefficient design," 2013 18th IEEE European Test Symposium (ETS), Avignon, 2013, pp. 1-6.
- [10] M. Shafique et al., "Invited Cross-layer approximate computing: from logic to architectures," in Proceedings of the 53rd Annual Design Automation Conference on - DAC '16, Austin, Texas, 2016, pp. 1-6.
- [11] H. Amrouch et al., "Towards Aging-Induced Approximations," in Proceedings of the 54th Annual Design Automation Conference 2017 on - DAC '17, Austin, TX, USA, 2017, pp. 1–6.
- [12] V. Gupta et al., "Low-Power Digital Signal Processing Using Approximate Adders," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 32, no. 1, pp. 124–137, Jan. 2013.
- [13] V. Gupta et al., "Low-Power Digital Signal Processing Using Approximate Adders," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 32, no. 1, pp. 124–137, Jan. 2013.
- [14] M. Shafique et al., "A low latency generic accuracy configurable adder," 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, 2015, pp. 1-6. doi: 10.1145/2744769.2744778
- [15] P. Kulkarni et al., "Trading Accuracy for Power with an Underdesigned Multiplier Architecture," 2011 24th Internatioal Conference on VLSI Design, Chennai, 2011, pp. 346-351. doi: 10.1109/VLSID.2011.51
- [16] S. Su et al., "A Novel Heuristic Search Method for Two-level Approximate Logic Synthesis," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., pp. 1–1, 2019.
- [17] Z. Zhou et al., "DALS: delay-driven approximate logic synthesis," in Proceedings of the International Conference on Computer-Aided Design - ICCAD '18, San Diego, California, 2018, pp. 1–7.
- [18] S. Guo et al., "Towards reliability-aware circuit design in nanoscale FinFET technology: – New-generation aging model and circuit reliability simulator," in 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Irvine, CA, 2017, pp. 780–785.
- [19] CliffordWolf.2016. Yosys open synthesis suite. (2016).
- [20] Berkeley Logic Synthesis and Verification Group. 2018. ABC: A system for sequential synthesis and verification, release 70330. (2018).
- [21] S.Yang. 1991. Logic synthesis and optimization benchmarks user guide: version 3.0. Microelectronics Center of North Carolina (MCNC).
- [22] H. Amrouch et al., "Estimating and Optimizing BTl Aging Effects: From Physics to CAD," 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Diego, CA, 2018, pp. 1-6.
- [23] A. Mishchenko *et al.*, "DAG-aware AIG rewriting: a fresh look at combinational logic synthesis," 2006 43rd ACM/IEEE Design Automation Conference, San Francisco, CA, 2006, pp. 532-535.