Characterizing Approximate Adders and Multipliers for Mitigating Aging and Temperature Degradations

Francisco J. H. Santiago, Honglan Jiang, *Member, IEEE,* Hussam Amrouch, *Member, IEEE,* Andreas Gerstlauer, *Senior Member, IEEE,* Leibo Liu, *Senior Member, IEEE,* and Jie Han, *Senior Member, IEEE*

***Abstract*—The performance of nanoscale semiconductor technologies has become susceptible to high temperatures and aging phenomena. While guard-bands have conventionally been used to combat degradation-induced timing violations, approximations have recently been leveraged to compensate for degradations in lieu of adding timing guard-bands, without a loss in performance. However, only simple approximation techniques such as truncation have been considered in prior work. In this paper, a wide range of approximate arithmetic circuits including adders and multipliers using various sophisticated approximation techniques are investigated to cope with aging- and temperature-induced degradations. To this end, approximate circuits are first characterized for their delay increase under degradations. With this, we then determine the approximation level required to compensate for guard-bands under different degradations. Degradation-aware logic synthesis results show that the simple use of truncated arithmetic circuits leads to a higher quality loss compared to using other approximate circuits. However, a truncated multiplier has the lowest error distance towards a reliable operation in 10 years. The approximate multipliers with configurable error recovery are most suitable when the level of degradation is higher, e.g., at a temperature of 70 °C. The characterization of degradation at the circuit level is then used for design exploration at the architecture level without the need for further gate-level simulations. For three different image processing applications, experimental results show that guard-bands can be mitigated while maintaining an output result with a high visual quality.**

***Index Terms*—Approximate computing, arithmetic circuits, performance and reliability, negative bias temperature instability.**

I. Introduction

SINCE the prediction of Gordon Moore in 1965, the scaling of transistors has brought significant gains in performance and power-efficiency [1]. However, transistors have also become more susceptible to various kinds of degradations, which affect the reliability of a circuit over its lifetime. These degradations, typically caused by process, voltage and temperature (PVT) variations, generate performance variability and thus jeopardize the correct operation of an integrated circuit. Specifically, device aging and temperature phenomena have become a real concern for the reliability of nanoscale CMOS devices (with a feature size $\leq$ 45 nm) [2].

Among these phenomena, the aging and temperature effects alter the threshold voltage and carrier mobility, which in turn leads to slower transistors. Catastrophic errors occur when the delay of the logic gates exceeds the clock period. Therefore, delay guard-bands are commonly added on top of the nominal critical path delay to avoid timing violation errors. Unfortunately, the circuits are typically over-designed with pessimistic guard-bands because it is hard to accurately analyze the worst-case scenarios, resulting in a significant loss in performance. Designers have proposed different methodologies to optimize the circuit performance while maintaining the reliability over its projected lifetime. However, those conventional techniques significantly increase the overall chip cost due to a higher area, power or a long simulation time. Since performance and energy efficiency have become critical aspects of emerging hardware architectures, a cost-effective design methodology that guarantees reliability without incurring hardware overhead is essential in the era of advanced CMOS technologies.

Approximate computing has recently emerged as a solution to improve reliability due to degradations without adding unnecessary hardware overhead. The principles of approximate computing have been explored in arithmetic circuits such as adders and multipliers to reduce area, energy consumption, and the critical path delay [3]. While computation errors are usually not desirable, approximate computing takes advantage of the fact that many applications, e.g., in image/signal processing or machine learning, may tolerate a certain level of errors as long as the results meet the quality requirement. Therefore, in error-tolerant applications, approximate circuits with a shorter critical path delay can be used in combination with conventional techniques to mitigate degradation effects by trading off output quality in lieu of adding delay guard-bands. That is, we can run the chip at the maximum clock frequency while maintaining a sufficient output quality rather than adding overhead to the chip for an unnecessarily optimized result.

In prior work [4], [5], [6], the idea of mitigating delay

This work is supported in part by the National Natural Science Foundation of China under Grant 62104127, in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Project RES0025211 and Project RES0048688, and in part by the German Research Foundation (DFG) grant number 428566201 (ACCROSS, AM 534/3-1).). (Corresponding authors: Honglan Jiang; Jie Han.)

F. J. H. Santiago was with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada. He is now with Intel, Zapopan 45017, Mexico (e-mail: fh@ualberta.ca).

H. Jiang is with the Department of Micro-Nano, Shanghai Jiao Tong University, Shanghai, 100084, China (e-mail: honglan@sjtu.edu.cn).

H. Amrouch is with the University of Stuttgart, Chair of Semiconductor Test and Reliability (STAR), Stuttgart, 70569, Germany (e-mail: amrouch@iti.uni-stuttgart.de).

A. Gerstlauer is with the Department of Electrical and Computer Engineering, the University of Texas at Austin, Austin, TX 78712, USA (e-mail: gerstl@ece.utexas.edu).

L. Liu is with the School of Integrated Circuits, Tsinghua University, Beijing, 100084, China (e-mail: liulb@tsinghua.edu.cn).

J. Han is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada (e-mail: jhan8@ualberta.ca).

guard-bands due to aging or temperature effects via truncation of the least significant bits (LSBs) in the inputs of an arithmetic circuit was introduced. However, while a wide variety of approximation schemes beyond simple truncation have been proposed [7], [8], no existing work has studied their effectiveness in mitigating delay guard-bands under aging and temperature degradations. Towards this goal, this paper represents a significant extension of prior work to make the following novel contributions:

1) Various approximate adders and multipliers are evaluated in terms of their delay behavior under different levels of degradations due to aging and temperature effects.
2) Since different logic exhibit different timing behaviors, we demonstrate how different levels of precision are required to sustain reliability when delay-optimized versus area-optimized design constraints are employed during the synthesis process.
3) An evaluation and comparison of approximate circuits is performed to determine which design provides the best performance towards trading-off guard-bands for a minimum quality loss in the output, evaluated against the accurate designs from the Synopsys DesignWare library.
4) A design-space exploration at the architectural level is introduced to determine optimal approximate component to instantiate without the need for running time-consuming simulations for applications that demand different qualities. Our simulation results show that truncation of LSBs is not the most effective scheme towards guard-band mitigation for high-performance or low-power applications.

The rest of this article is organized as follows. Section II discusses the sources of transistor variations and degradations, and summarizes the most recent work to characterize and/or mitigate delay guard-bands in digital circuits. In Section III, the framework to pre-characterize approximate circuit libraries towards guard-band reduction is presented at the circuit level. Section IV discusses how by bringing degradation aware at earlier stages, we could optimize the architectural level for different hardware requirements. Section V and VI introduces the experiments and results for the characterizations on circuit and architecture levels, respectively. The impact of variation on circuit performance under different levels of degradations are discussed. Also, the reductions of precision in approximate circuits are characterized to guard-band the accurate designs under aging and temperature effects. Finally, Section VII concludes this work.

II. Background

This section presents the sources of performance variation, aging and temperature phenomena in particular, and introduces the state-of-the-art design methodologies that have previously been proposed to improve circuit reliability.

A. Sources of Transistor Variations and Degradations

Historically, semiconductor manufacturers used to employ the same scaling factor for the supply voltage (S_V) and transistor length (S_L) to increase performance and maintain the electric field constant. Unfortunately, the simple scaling of transistor feature sizes appears to have broken down. This occurs because the supply voltage cannot be scaled down anymore, or it would fall below the threshold voltage. With the continuous search to increase performance, the rule of scaling in semiconductor fabrication changes from ($S_V = S_L$) to ($S_L < S_V$) [9]. As a result, the electric field across the channel and gate has been increasing in the most recent semiconductor technologies. With a higher electric field, the aging phenomena become stronger, which in turn, may break the gate dielectric. This phenomenon, called time-dependent dielectric breakdown (TDDB) may result in the total failure of a circuit. To solve this problem, semiconductor manufacturers replaced the material employed to build the gate dielectric layer with a more resistant material (i.e. the high-k dielectric material) [10]. Despite the probability that a TDDB occurs has decreased, the employment of such new materials impacts the transistor characteristics due to other aging phenomena [9].

The most prominent degradations in transistors due to the use of high-k materials has been categorized as bias temperature instability (BTI) and hot carrier induced degradation (HCID). BTI and HCID were reported since 1966, but they only became a significant issue once the gate oxide thickness is scaled to values lower than 1.5 *nm*. Since a high electric field is the key source behind aging induced degradations, carriers, which are accelerated by the electric field, collide with the gate rather than moving between the drain and the source when a pMOS/nMOS transistor is in operation. The collision degrades transistors due to the charges that get trapped inside the dielectric. While a vertical electric field over the gate stimulates the BTI, a lateral electric field across the channel stimulates the HCID. Both phenomena significantly degrade the carrier mobility and threshold voltage in the transistors, thus reducing performance over the transistor's lifetime [9].

Temperature fluctuations are caused by elevated ambient temperature or heat due to power dissipation across a chip. Temperature imposes a design constraint called thermal design power that determines the maximum amount of heat that the cooling system can dissipate [11]. Once the temperature starts rising beyond the limits, the transistor becomes slower due to reduced carrier mobility and interconnect resistance, which in turn leads to timing violations [12].

B. Techniques for Coping with Degradations

Timing guard-bands have been commonly used in the past years for coping with degradations. As transistors become slow due to the aforementioned degradations, designers can increase the clock period in such a way that the critical path delay would be smaller than the clock period during a circuit's expected lifetime. Unfortunately, this leads to a significant loss in performance. Alternatively, a voltage guard-band can be added on top of the nominal voltage to increase the transistor's current and thus the switching speed [13] [14] [15]. This approach allows us to run the circuit at the highest performance, but it will also consume more power. Therefore, in this article, we focus primarily on timing guard-bands.

The main drawback of timing guard-bands is that they require analysis of the worst-case scenarios. However, worst-case conditions are hard to define. On the one hand, it is unknown if the transistor's parameters will be shifted during the process of synthesis, layout, or integration of the whole chip. On the other hand, it has been recently demonstrated that the worst-case degradation uniformly applied to each transistor does not capture the actual worst-case of a cell [16]. Therefore, the circuits are typically over-designed with pessimistic guard-bands for conditions that will rarely happen. Based on these observations, more sophisticated techniques have been investigated to mitigate the impact of guard-bands in integrated circuits including, for example, design-time synthesis, adaptive techniques and, most recently, approximate computing.

1) Design-time Synthesis: To accurately contain/reduce guard-bands, authors in [17] generate cell libraries to model aging-induced degradations. These libraries can be employed during the process of synthesis and timing analysis with existing commercial tools. The authors demonstrated that ignoring carrier mobility leads to overestimating guard-bands by almost 20%. This work is consequently improved in [18] to characterize the impact of aging on dynamic and static power. The same authors also proposed static and adaptive optimization techniques for temperature guard-bands in [19]. Both have been evaluated in five well-known processors, and experimental results show that delay guard-bands can be reduced by 22% compared with traditional approaches.

2) Adaptive Techniques: With the increasing demand for high-speed circuits, a permanent loss in performance even when degradations have not yet occurred, is not acceptable. This fact has increased interest in adaptive techniques that maintain performance while ensuring reliability. For instance, Sadi *et al*. present a new framework with self-adaption capability against aging-induced degradation [20]. This methodology uses built-in self test (BIST) to monitor critical paths in a design. Therefore, the primary step in this framework was the introduction of automatic test pattern generation (ATPG) targeting the high-usage critical paths under aging-induced delay. The BIST mechanism feeds a machine learning algorithm with the results to predict the aging degradation state. Predicted results are used to activate a remedy against timing degradation.

3) Approximate Computing: Most recently, approximate computing has been employed to mitigate guard-bands at the circuit level. Using *aging-induced approximation* [4], Amrouch *et al*. show that aging-induced timing errors lead to an unacceptable quality drop even for inherently error-tolerant applications. To address this problem, instead of using delay guard-bands to sustain reliability, the authors converted aging-induced timing errors into controllable and deterministic errors coming solely from approximating and hence reducing the critical path delay of arithmetic computations. Experimental results show that truncation of 10 and 3 LSBs is enough to sustain reliability in a 32-bit adder and multiplier, respectively. In the context of an image processing application, this methodology not only eliminated delay guard-bands at the architectural level but also enhances energy efficiency by 13% with a peak signal-to-noise ratio (PSNR) higher than 30 dB.

Boroujerdian *et al*. proposed two approaches for synthesizing delay-configurable circuits to overcome temperature variations and consequently narrow guard-bands [5]. These configurable circuits minimize quality losses by dynamically and adaptively applying quality scaling in the presence of temporary circuit degradations. This approach also takes advantage of automatic re-partitioning algorithms from the EDA tools. The first approach consists of duplicating approximate arithmetic circuits to overcome different levels of temperature without sharing any resources among them. This approach accurately sets the narrowest guard-bands for each possible scenario, but suffers from significant energy and area overhead. The second approach decreases the energy and area overhead by sharing the resources among the circuits. Selection of the appropriate approximate circuits is based on the measurement of temperature in real time. Results of an IDCT application show up to a 21% speedup with a PSNR higher than 39 dB in the output image.

An adaptive technique was presented in [6] to use truncation-based approximation for the compensation of aging-induced degradation. This design measures the effects of aging using a monitoring system in real time. If the result does not meet the timing requirement, the proposed design excludes the computation of the LSBs instead of adjusting the values of voltage or clock frequency. By this means, the designers guarantee that the clock period is large enough to perform the computations in the most significant bits (MSBs). Experimental results, compared to conventional guard-bands approaches, show an improvement of 21.45% and 10.78% for dynamic and static power, respectively.

C. Discussion

To cope with the degradations due to aging and temperature effects, a large number of methodologies have been proposed to improve hardware efficiency and prolong chip lifetime. However, most of the methods suffer from energy and delay overhead. On the other hand, approximate computing has emerged as another solution to mitigate degradations in lieu of adding delay guard-bands. However, only the truncation of LSBs has been investigated so far to overcome degradations, which cannot show the effectiveness of approximate computing. Considering that a multitude of dedicated designed approximate arithmetic circuits have been proposed and studied recently [21], all these studied circuits have not been evaluated in the context of reliability and their applicability towards compensating for degradations. Therefore, a comparative study needs to be performed in order to determine if there are more effective approximation schemes beyond just simple truncation with better trade-offs among error, power, and speed.

III. Approximate Circuit Characterization under Aging and Temperature Effects

Instead of using timing guard-bands to guarantee reliability, our methodology consists of converting degradations to deterministic and controllable errors coming solely from an approximate arithmetic circuit. Specifically, an approximate circuit is characterized by considering possible degradations, to obtain

the same effective performance as an accurate circuit without any degradation. This can be understood as maintaining the performance of an accurate arithmetic component at the cost of a quality loss in the output. By this means, no guard-band is required at the component level to maintain reliability during its lifetime.

The methodology to trade-off degradations for precision is summarized in Fig. 1. Different from [4], we present a comprehensive framework using the main characteristics of approximate computing. The most important steps are summarized below.

Logic Synthesis: To achieve the optimal performance of each approximate circuit under different design constraints, accurate sub-adders are coded in the hardware description language (HDL) with a high-level description (identified using "+") and used as building blocks in the approximate component. For instance, the final addition for most of the approximate multipliers at the partial product accumulation stage uses the "+" operation. This allows re-using the HDL during the synthesis process when we aim high-performance or low-power synthesis. By employing this coding strategy during the design phase and the "compile_ultra" option in the Design Compiler tool, we can obtain a higher quality of results using aggressive and efficient optimization algorithms from the Synopsys tools instead of exploring approximate circuits with different adder tree topologies [22].

Verifying Timing Across Degradations: An exhaustive timing verification of all possible scenarios is impractical, and so the use of STA with Synopsys PrimeTime is used to verify the required timing margin to overcome different degradations in the optimized netlist [23]. Timing model libraries are required to provide detailed information about the cells under different stress conditions. While some technology libraries include timing models of PVT variations, other timing models that describe how transistors degrade considering aging or temperature are publicly available in [24]. The output of this stage will be an estimation of the required time to avoid violations in the critical path when the chip is degraded.

Timing Goal: The objective is to accurately trade off degradations of an accurate circuit for a quality loss, rather than using delay guard-bands to overcome the degradations. Therefore, if the delay of the approximate circuit under respective degradations is larger than the delay of the timing goal (set for the accurate circuit without degradations), then the whole process is repeated by reducing the precision in the approximate circuit. We assumed that gains in performance are obtained each time we reduce the precision in the approximate circuits.

Obtaining Error Metrics: Different error metrics such as the error rate (ER), the normalized mean error distance (NMED), the mean relative error distance (MRED), and the mean squared error (MSE) have been employed to quantify the accuracy of the approximate circuits [25]. The ER is defined as the percentage of erroneous outputs among all outputs, the error distance (ED) as the absolute distance between the approximate and the accurate result, and the mean error distance (MED) as the mean of all possible EDs. The definition of NMED, RED, MRED, and MSE are given below.

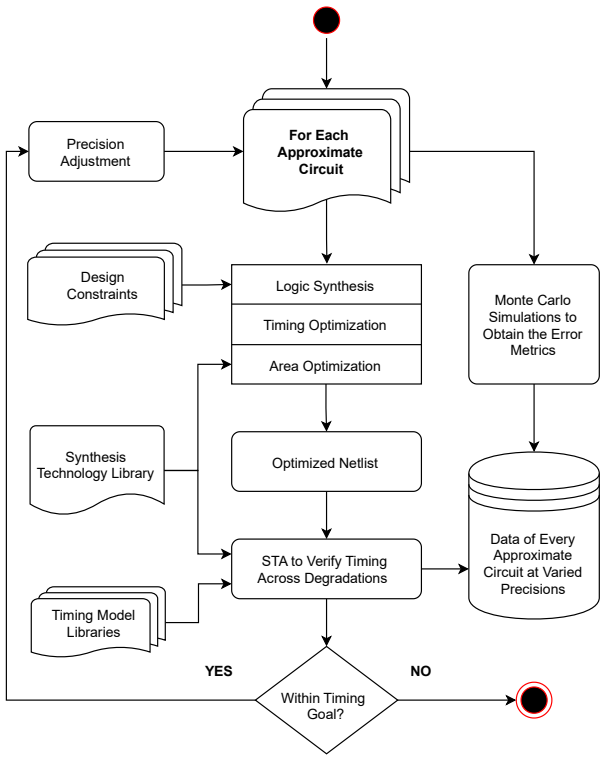


Fig. 1. Design methodology at the circuit level to convert degradations to controllable errors using approximate circuits (adapted from [4])

$$\text{NMED} = \frac{\text{MED}}{MAX}, \tag{1}$$

$$\text{RED} = \left| \frac{\hat{y} - y}{y} \right|, \tag{2}$$

$$\text{MRED} = \frac{\sum_{i=1}^{N} (\text{RED})}{N} \tag{3}$$

and

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{4}$$

where y, $\hat{y}$, N and MAX denote the accurate result, the approximate result, the total number of possible input combinations, and the maximum magnitude of the output of an accurate design, respectively.

The selection of the right metrics is a key step during the process of evaluation at the circuit level. For instance, an arithmetic error metric (e.g., MRED or MSE) would often be more useful than the ER to evaluate the impact on a target application. On the other hand, obtaining several error metrics with all possible input combinations may be overly time-consuming and computationally expensive. As a practical strategy, Monte Carlo simulations are employed to evaluate the functionality of each approximate circuit design. This statistical technique applies a randomly selected subset of the set of all possible input vectors based on certain probability distributions (e.g., uniform, Poisson, Gaussian, or exponential). In the context of this article, we employed 10 million uniformly distributed random input combinations to evaluate

16-bit approximate multipliers and adders. The error metrics obtained with the Monte Carlos simulations are stored in a database for further comparison at the architectural level (see Section IV).

IV. Architectural Exploration

At the architectural level, timing plays the most crucial role as the design progresses through the design flow. Tightened requirements considering degradations further increase the complexity of meeting the delay criteria. In the worst-case scenarios, violations in timing lead to adding unexpected delay guard-bands or architectural modifications in the RTL, which significantly affect the chip cost. Therefore, we discuss in this section how, by bringing awareness of aging or temperature effects using the degradation-aware approximate libraries (see section V-A) during the design phase, we can efficiently synthesize a design that meets all the timing requirements and dissipates low power for the entire design. Furthermore, because different approximate techniques provide different trade-offs, the proposed methodology using our libraries enables design-exploration without the need for running time-consuming simulations.

The process for converting degradations into controllable errors for an architectural-level design can be divided into multiple stages, as shown in Fig. 2. In this section, the design methodology is extensively applied to overcome aging and temperature degradations in the computing part. However, this methodology can also be applied to reduce the significant increase of delay in the die due to the place-and-route stage or On-Chip Variations (OCV) [23]. This methodology has been taken and improved from [4] to optimize the final result by means of a design exploration to select the best approximate units pre-characterized as described in Section III. In the following, we explain in detail the methodology step by step.

Obtaining Timing Constraints: First, the architecture is synthesized to obtain the critical path (CP) delay in the absence of any degradation ($t_{cp}(freshDesign)$). This delay represents the required timing constraint that the whole design must fulfill under the targeted aging or temperature stress condition. It is assumed that the critical path belongs to the arithmetic circuits with the purpose of introducing approximations in the computations, and thus, reduce the critical path delay. Other components, such as control units, can be protected through traditional techniques, such as using stronger gates [13].

Estimating degradations: Under a specific level of stress, STA is performed for the whole design to obtain the delay of every combinational datapath block (Bk) within the netlist ($t_{Bk}(postStress)$). This allows us to calculate the available timing slack $t_{Bk}(slack)$ (see (5)) between the timing constraint and the delay of each block considering degradations ($t_{Bk}(postStress)$). While a *positive* time slack means no guard-bands are needed, a *negative* value (i.e., $t_{Bk}(slack) < 0$) means that timing violations will occur in the corresponding component. Hence, delay guard-bands are required to avoid catastrophic errors, which leads to hardware performance

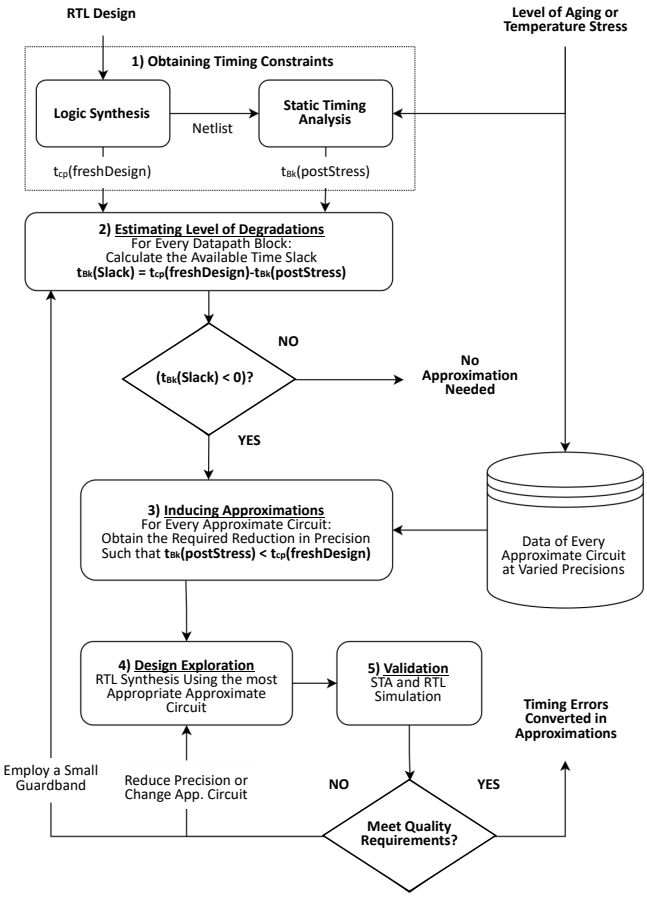


Fig. 2. Design methodology at the architectural level to convert degradations to controllable errors using approximate circuits (adapted from [4]).

(speed) loss. Instead, the negative slack can be compensated with approximations to maintain the speed in the architecture.

$$t_{Bk}(slack) = t_{cp}(freshDesign) - t_{Bk}(postStress) \quad (5)$$

Inducing Controllable Approximations: Considering every block Bk contains an arithmetic circuit that can be approximated, the characterized components in Section V-B (see Fig. 1) can be employed to efficiently compensate for the existing negative time slack according to the target application. Depending on how large the existing time slack is, the precision reduction can be the maximum precision reduction allowed for the approximate circuit or smaller.

Design Exploration: Different approximation schemes with distinct circuit characteristics may require a different level of precision to meet the same timing goal. Therefore, at this stage, an approximate circuit is selected according to the output quality or circuit characteristics. Specifically, the approximate circuits obtained from the previous step are compared in terms of accuracy and hardware overhead. At this step, the first priority is to ensure the accuracy of the application. For the approximate designs satisfying the accuracy constraint, the most efficient design with the smallest power dissipation or critical path delay is chosen, depending on the application requirements (e.g., low-power or high-performance). In this case, the most appropriate approximate circuits are found. To assess

the accuracy of a complex application with recurrent arithmetic operations, several general input datasets are sufficient for the stimulus. However, for a simple architecture with single layer of arithmetic operation, Monte Carlo approach should be used to show the accuracy of the approximate designs. After determining the most appropriate approximation, we implement corresponding modifications in the RTL and repeat the process of synthesis to optimize the glue logic surrounding the approximate components. By this means, we are also trading power and speed while keeping the reliability high.

Validating Timing Constraints and Quality Output: We then perform degradation-aware STA with the new netlist and a functional RTL simulation to ensure that we meet the timing constraints and output quality, respectively. Our methodology is independent of the specific constraints and quality metric chosen for exploration. The utilized quality metrics and constraints depend on the type and requirements of applications. Also, several levels of quality loss can be set as accuracy constraints to achieve different hardware gains. Note that there is a small likelihood that a small negative timing slack remains. This can be due to an increase in the degradation-induced delay in the glue logic surrounding components. In such a case, another reduction of precision to compensate for the remaining slack will be necessary. As a second option, another approximate circuit with a larger error can be investigated (by a design exploration). If the final quality output is not sufficient, the precision can be increased at the cost of a small guard-band. However, such a guard-band will be significantly smaller than the original one when no approximations are applied.

V. Circuit Characterization Results

In the following, we first characterize basic arithmetic circuits, i.e. adders and multipliers, under aging and temperature effects. Specifically, we measure the degradations in the critical path delay of various approximate arithmetic circuits. The approximations of different designs are then obtained for the performance degradations without using guard-bands.

A. Adder and Multiplier Degradations

Similar to conventional accurate circuits, approximate circuits require delay guard-bands to overcome the aging or temperature effects. Otherwise, non-deterministic errors will occur during their lifetime. Therefore, in the following we investigate how the performance of an approximate circuit varies by accurately estimating the required guard-band to overcome degradations in each circuit. This is of special interest since recent literature has shown that a critical path in a circuit may not continue to be critical under different levels of workload activity or voltage [26]. Similarly, the work in [19] confirmed that the delay of some circuits increases by 70%, while others may only be affected by 10% when the temperature rises from a typical value (e.g., $25^{\circ}C$) to the worst-case value ($70^{\circ}C$). In particular, we are interested in an investigation to find out if the performance of the approximate circuits may be different considering the required delay guard-bands to overcome degradations due to aging and temperature effects.

We employ the design methodology presented in Fig. 1, but consider all possible levels of precision of each approximate circuit for comparison purposes. Logical representations of 16-bit approximate adders and multipliers are implemented using Verilog and VHDL. The Synopsys Design Compiler was employed in the process of synthesis using a 45-nm Nangate process technology [27]. For a fair comparison, all designs were synthesized with the same timing constraint. We synthesized designs with a constraint of zero. The "ultra compile" option is also used during the synthesis process to maximize the quality of the results. The degradation-aware cell libraries from [24] are used to characterize the aging and temperature effects in the approximate circuits. The accuracy of the approximate designs is evaluated with Matlab through Monte Carlo simulations. 10 million random inputs with normal distributions were employed to obtain the error metrics. As mentioned before, this experiment is for evaluating the impact on delay of the aging and temperature effects; however, an extended comparison considering remaining errors and hardware metrics under the nominal delay is discussed in [28].

Approximate circuits can be divided into traditional manual and automatic designs. Manual designs are often obtained by purposefully modifying the accurate circuits, whereas an automated method employs advanced algorithms to randomly and iteratively remove some logic gates until the design requirements are met. See [21] for a comprehensive evaluation and comparison of approximate arithmetic designs using various approximation techniques. Note that our proposed approach is general and can be applied to any approximate circuit. In this paper, several representative approximate adders and multipliers for each approximation technique are considered for characterization under aging and temperature effects. Table I shows the definitions of the considered designs. We expect insights from our study to transfer similarly to other approximate designs. Note that a type of hardware-efficient approximate multipliers using logarithmic approximation [29], [30] are not considered here due to their relatively high errors. The MSE of the 16-bit approximate logarithmic multipliers is at least at the level of 10^{15}. Also, the Booth multipliers based on approximate encoding [31], [32] are not studied because they generally show longer delay than the truncated Booth multiplier (TBM) for a similar MSE.

Figs. 3 and 4 show the large design flexibility and accuracy range of the approximate adders and multipliers, respectively. Note how the performance (delay) varies according to the levels of stress. Although we found cases where the critical path in an approximate circuit does not remain as the critical path after transistor degradations, the results indicate that their relative performance in terms of error remains the same under different levels of degradations. For instance, the CGPAs show the highest performance when we aim for an approximate circuit with a low MSE regardless of the level of aging or temperature degradation.

Specifically, CGPAs are the fastest circuits for an MSE of up to 10^2 independently of the design constraint. At a larger error, the LOA, ESA and CSPA are the most efficient at high-speed, respectively. With regards to the multipliers, CGPMs show the best performance at a given low MSE. Nevertheless, TAM2

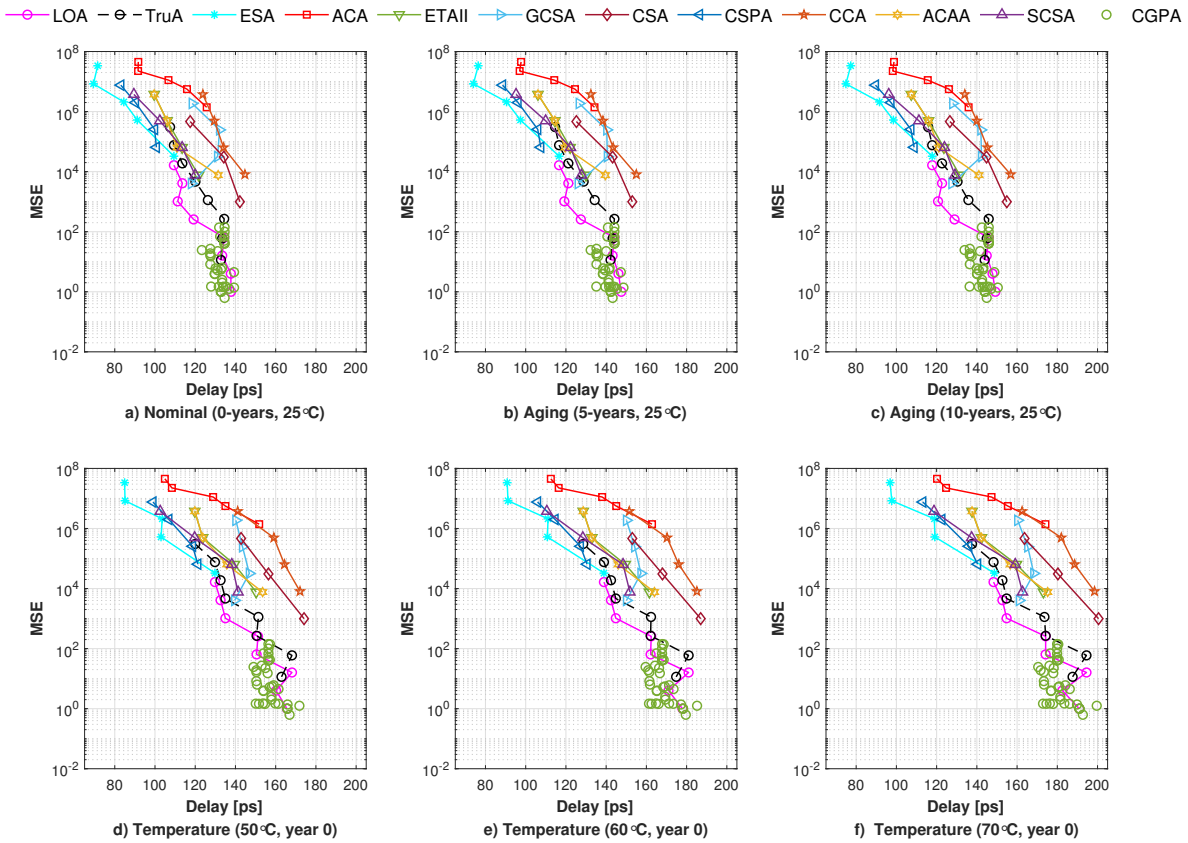


Fig. 3. Characterizing delay-error tradeoffs of 16-bit high-performance approximate adders under different aging and temperature effects. The parameter k for LOA and TruA ranges from 2 to 9, for ESA and ACA from 8 down to 3 (except 7), CSA from 5 down to 3 and for the remaining adders from 6 down to 3, all from right to left. Regarding the CGPAs, the configurations with the lowest error metrics for a specific delay are reported.

TABLE I
DEFINITION OF APPROXIMATE ARITHMETIC CIRCUITS

Adder	**Definition**
TruA	Truncated adder
LOA	Lower-part-OR adder [8]
CGPA	Cartesian genetic programming-generated adders [33]
ETAII	Error-tolerant adder type II [34]
SCSA	Speculative carry selection adder [35]
ESA	Equal segmentation adder [36]
ACAA	Accuracy-configurable approximate adder [37]
CSPA	Carry speculative adder [38]
CSA	Carry skip adder [39]
ACA	Almost correct adder [40]
GCSA	Generate signals-exploited carry speculation adder [41]
CCA	Consistent carry approximate adder [42]
Multiplier	**Definition**
TruM	Truncated multiplier
PPAM	Partial product perforation-based multiplier [43]
AM 1/2	Approximate multiplier with configurable error recovery 1/2 [7]
TAM 1/2	AM 1/2 with half truncated partial products 1/2 [7]
CGPM	Cartesian genetic programming-generated multipliers [44]
ACM	Approximate compressor-based multiplier [45]
ICM	Inaccurate counter-based multiplier [46]
UDM	Underdesigned multiplier [47]
TBM	Truncated modified Booth multiplier truncating two inputs
TBMS	Truncated modified Booth multiplier truncating a single input

and TAM1 become faster for a larger MSE. For the signed multipliers, the TBM shows a higher speed than TBMS for a large MSE, whereas the TBMS can be faster for a small MSE. Similar results for both approximate adders and multipliers were observed considering the other error metrics such as the MRED and NMED.

B. Degradation-Induced Approximations

We applied the methodology proposed in Fig. 1 to characterize the required level of 16-bit approximate adders and multipliers towards aging and temperature degradations. In the process, we analyzed our results to determine which approximate design overcomes degradations with the minimum error possible. The 45-nm NanGate process technology with a supply voltage of $1.2V$ is employed during the process of synthesis [27]. In these experiments, we characterize approximate circuits towards aging- and temperature-induced approximation using two different design constraints: high-performance and low-power. The degradation-aware cell libraries are employed during the STA with PrimeTime to accurately estimate the effects of aging and temperature in the circuits [19] [24]. To be competitive with state-of-the-art methods, the timing goal is defined as the clock frequency given by the accurate circuit from the Synopsys DesignWare library in the absence of degradations. Finally, the MSE, which is correlated with

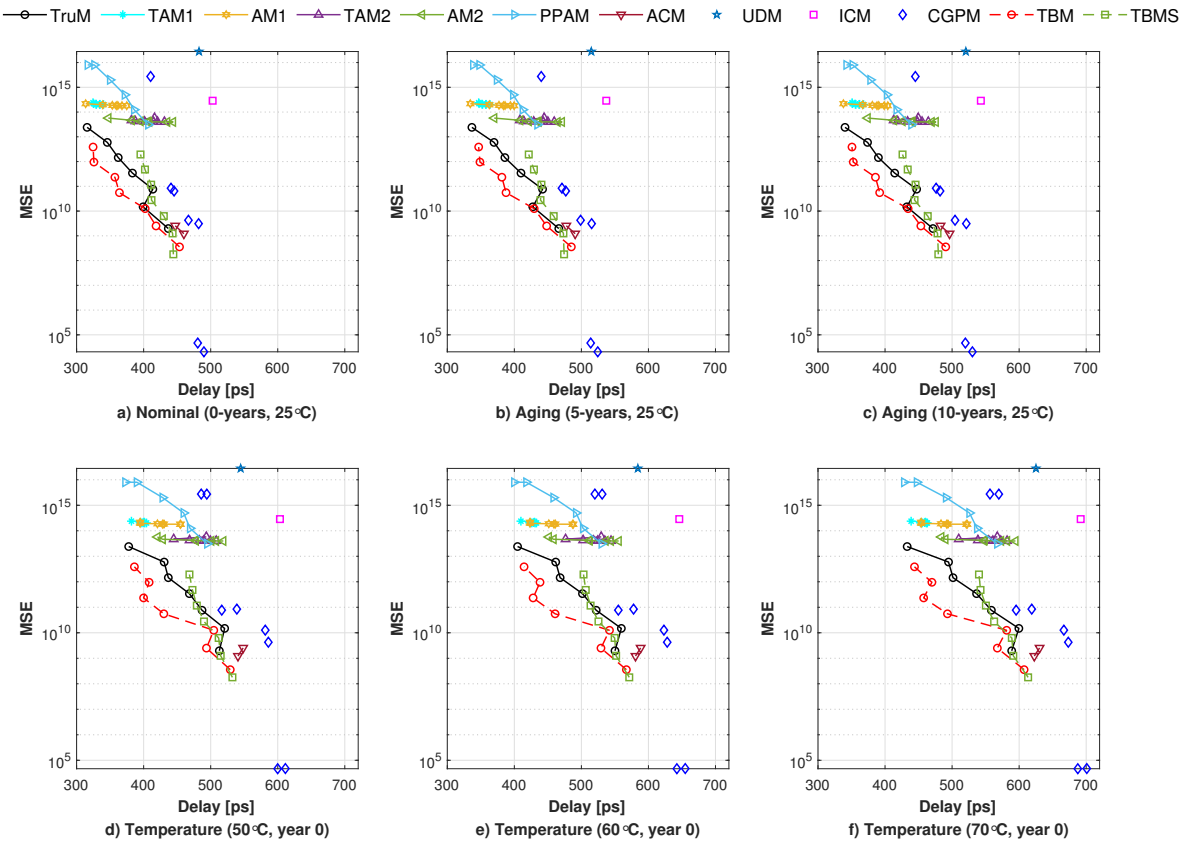


Fig. 4. Characterizing delay-error tradeoffs of 16-bit high-performance approximate multipliers under different aging and temperature effects. The number of truncated LSBs for TruM, TBM, and TBMS is from 1 to 7 from right to left. The number of MSBs used for error compensation is from 16 down to 10 for AM1, AM2, TAM1, and TAM2 from right to left. The mode number for ACM is from 4 to 3 from right to left. Regarding PPAM and CGPMs, the best designs are reported.

the PSNR, is used for ranking the approximate circuits and determining the best design. As mentioned, the objective is to maintain the performance of an accurate arithmetic component at the cost of a quality loss, rather than using delay guard-bands.

Approximation Towards Aging Degradation: Tables II and III show the required level of precision of each approximate circuit to achieve our timing goal. Note that the worst-case aging-induced delay (in 10 years) for the approximate circuits is lower than the delay of the accurate circuit in the absence of aging (0 year). Therefore, we can use any of these approximate circuits instead of the accurate one with the assurance that delay guard-bands would not be required to ensure a circuit meeting the timing goal after 10 years.

As can be seen for the adders synthesized for high-performance (see Table II), CGPA-156 is the best design towards aging-induced approximation. The CGPAs do not only show the lowest MSE, but also the lowest MRED among all the approximate adders. Considering only the manual designs, LOA-6 and truncated adder (TruA)-5 are the most effective circuits. On the other hand, if ER is the most important measure for a target application, the ACA, ETAII and GCSA (with an ER less than 1%) considerably outperform CGPA-156, LOA-6 and TruA-5. The use of these speculative adders (ETAII-6 and GCSA-3) is highly recommended in applications where the probability of a large carry chain is relatively small.

Regarding the high-performance multipliers, the TruM-2 shows the best performance with the lowest values in the MSE, MRED and ER. The results indicate that the truncation of 2 bits is enough to ensure the target reliability for 10 years. The circuit of the CGPM-4A19[1] is not as efficient as the manual designs toward aging-induced approximation. The CGPMs were originally designed to simultaneously improve power and delay (or power-delay product (PDP)), which limits the gains in speed. In terms of signed multipliers, experimental results show that TBM-2, with lower errors, is more effective than TBMS.

Table III shows the results towards-aging induced approximation for the circuits synthesized for low power. Different from the high-performance circuits, we observed that automatically generated designs outperform manual designs. Regarding the approximate adders, the CGPA-449 has the lowest MRED and MSE values, followed by LOA-3 and TruA-2. The least effective designs in terms of MSE are the CSPA-6 and ACA-8. Regarding the approximate unsigned multipliers,

[1]4A19 represents the architecture A4 with the circuit 19 following the nomenclature in [44].

TABLE II
CHARACTERIZING 16-BIT HIGH-PERFORMANCE APPROXIMATE CIRCUITS TOWARDS AGING-INDUCED APPROXIMATION. THE CIRCUITS ARE RANKED BY THE MSE FROM THE LOWEST TO THE HIGHEST VALUE.

Component	Precision	Delay (*ps*) 0 year	Delay (*ps*) 10 years	ER (%)	MRED (10^{-3})	MSE
Adder	Accurate	138.6	149.5	0.00	0.00	0.00E+00
	CGPA-156	127.9	136.4	43.74	**0.16**	**1.49E+00**
	LOA-6	126.3	135.8	82.21	0.25	2.56E+02
	TruA-5	126.3	135.8	99.90	0.70	1.13E+03
	ETAII-6	121.7	131.5	0.73	**0.16**	7.63E+03
	SCSA-6	120.0	129.1	0.72	**0.16**	7.64E+03
	ESA-8	109.4	117.8	49.83	2.70	3.26E+04
	ACAA-5	111.0	121.1	2.29	0.65	6.34E+04
	CSPA-6	100.7	108.8	9.13	1.30	6.46E+04
	CSA-3	117.5	126.6	1.76	1.30	4.64E+05
	ACA-8	125.7	135.9	**0.68**	1.20	1.39E+06
	GCSA-3	118.5	128.0	10.76	4.00	1.86E+06
	CCA-3	123.8	134.0	23.66	7.90	3.73E+06
Unsigned multipliers	Accurate	457.7	489.6	0.00	0.00	0.00E+00
	TruM-2	399.2	432.5	**93.75**	**0.53**	**1.48E+10**
	PPAM-J0K8	406.9	439.3	99.61	14.37	3.10E+13
	AM2-14	408.3	441.0	99.10	2.32	4.02E+13
	TAM2-16	414.7	448.3	99.99	2.88	4.02E+13
	AM1-16	374.8	404.1	98.22	3.37	1.80E+14
	TAM1-16	334.4	361.3	99.99	6.45	2.02E+14
	CGPM-4A19	399.0	429.5	99.99	82.16	1.89E+15
Signed multipliers	Accurate	462.9	500.2	0.00	0.00	0.00E+00
	TBM-2	418.5	453.4	93.74	0.98	2.50E+09
	TBMS-4	411.7	444.0	93.74	2.51	2.77E+10

TABLE III
CHARACTERIZING 16-BIT LOW-POWER APPROXIMATE CIRCUITS TOWARDS AGING-INDUCED APPROXIMATION. THE CIRCUITS ARE RANKED BY THE MSE FROM THE LOWEST TO THE HIGHEST VALUE.

Component	Precision	Delay (*ps*) 0 year	Delay (*ps*) 10 years	ER (%)	MRED (10^{-3})	MSE
Adder	Accurate	942.9	1023.3	0.00	0.00	0.00E+00
	CGPA-449	821.7	916.9	**34.37**	**0.01**	**2.25E+00**
	LOA-3	821.7	891.8	57.80	0.03	4.00E+00
	TruA-2	821.7	891.8	93.74	0.06	1.15E+01
	CSA-5	427.0	457.4	0.02	0.01	9.93E+02
	GCSA-6	483.7	523.2	0.74	0.06	4.02E+03
	ETAII-6	600.0	643.3	0.73	0.16	7.63E+03
	SCSA-6	414.6	447.6	0.72	0.16	7.64E+03
	ACAA-6	532.4	577.3	0.72	0.15	7.64E+03
	CCA-6	382.8	413.0	1.50	0.12	8.05E+03
	ESA-8	458.4	497.1	49.83	2.70	3.26E+04
	CSPA-6	303.5	326.4	9.13	1.30	6.46E+04
	ACA-8	341.4	365.8	0.68	1.20	1.39E+06
Unsigned multiplier	Accurate	2009.3	2177.0	0.00	0.00	0.00E+00
	CGPM-1A222	1665.4	1802.1	**0.03**	**0.01**	**1.92E+05**
	TrunM-2	1757.0	1903.3	93.75	0.53	1.48E+10
	AM2-16	1560.1	1693.0	97.95	1.35	3.97E+13
	TAM2-16	1459.2	1580.2	99.99	2.88	4.02E+13
	PPAM-J0K9	1811.7	1982.1	99.80	26.08	1.24E+14
	AM1-16	1496.1	1622.6	98.22	3.37	1.80E+14
	TAM1-16	1336.9	1441.0	99.99	6.45	2.02E+14
Signed multipliers	Accurate	2059.7	2234.3	0.00	0.00	0.00E+00
	TBM-2	1815.2	1969.0	93.74	**0.98**	**2.50E+09**
	TBMS-3	1863.5	2022.1	**87.48**	1.18	6.26E+09

TABLE IV
CHARACTERIZING 16-BIT HIGH-PERFORMANCE APPROXIMATE CIRCUITS TOWARDS TEMPERATURE-INDUCED APPROXIMATION. THE CIRCUITS ARE RANKED BY THE MSE FROM THE LOWEST TO THE HIGHEST VALUE.

Component	Precision	Delay (*ps*) 25 °C	Delay (*ps*) 70 °C	ER (%)	MRED (10^{-3})	MSE
Adder	Accurate	143.5	196.7	0.00	0.00	0.00E+00
	LOA-10	100.1	137.6	94.36	4.00	**6.55E+04**
	CSPA-5	99.2	136.0	**11.31**	**2.70**	2.55E+05
	TruA-9	100.1	137.6	99.99	10.70	3.04E+05
	ESA-6	86.3	119.1	73.04	10.60	5.23E+05
	ACAA-3	99.7	137.6	18.86	10.30	3.73E+06
	ETAII-3	99.7	137.6	18.91	10.30	3.73E+06
	SCSA-3	85.4	118.7	18.89	10.20	3.73E+06
	ACA-4	90.4	124.7	16.65	18.90	2.24E+07
Unsigned multipliers	Accurate	467.4	643.8	0.00	0.00	0.00E+00
	TruM-7	315.9	433.0	99.99	15.57	**2.40E+13**
	AM1-11	330.5	453.9	**99.59**	9.85	1.97E+14
	TAM1-16	335.1	461.7	99.98	**6.45**	2.02E+14
	PPAM-J1k11	326.5	351.7	99.95	144.20	8.00E+15
Signed multipliers	Accurate	460.0	632.0	0.00	0.00	0.00E+00
	TBM-7	322.7	443.9	99.99	39.23	3.86E+12

we observed that CGPM-1A222[2] outperforms TruA-2 with a substantial difference in the MSE. AM1-16 and TAM1-16 are the least effective unsigned multipliers in terms of error; however, the gains in speed are much larger than the other circuits. For the signed multipliers synthesized under low-power optimization, the TBM-2 is faster than TBMS-3, with smaller ER, MRED and MSE.

Approximation Towards Temperature Effects: Tables IV and V show the required level of precision of each approximate circuit towards temperature-induced approximation. Note that the worst-case temperature delay (at 70 °C) for the approximate circuits is lower than the delay of the accurate design in a nominal temperature (25 °C). We also used the criteria of the MSE to rank the approximate circuits in the tables.

Table IV shows the results for the high-performance circuits. Different from the aging-induced approximation, we observed that some approximate design methodologies do not meet the timing goal, which means that those circuits still require a delay guard-band on the top of the maximum clock frequency (determined by the accurate circuit) to guarantee a deterministic function. As can be seen, automatically generated designs (CGPAs and CGPMs) are not present due to timing violations. Regarding the manual designs, LOA-10 shows the lowest MSE towards temperature-induced approximation at 70°C. CSPA-5 shows a larger MSE than LOA-10, but the MRED and ER are lower. In terms of approximate unsigned multipliers, the truncation of 7 bits (TruM-7) results in the smallest MSE followed by AM1-11. Although TAM1-16 shows a relatively large MSE, this approximate design has the lowest MRED. Regarding the signed multipliers, the truncation of 7 bits in the Booth multiplier (TBM-7) provides the best trade-off for degradation and performance compared to the TBMS (not present in the table).

[2]1A222 represents the architecture A1 with the circuit 222 following the nomenclature in [44].

TABLE V
CHARACTERIZING 16-BIT LOW-POWER APPROXIMATE CIRCUITS TOWARDS TEMPERATURE-INDUCED APPROXIMATION. THE CIRCUITS ARE RANKED BY THE MSE FROM THE LOWEST TO THE HIGHEST VALUE.

Component	Precision	Delay (*ps*) 25 °C	Delay (*ps*) 70 °C	ER (%)	MRED (10^{-3})	MSE
Adder	Accurate	942.8	1303.4	0.00	0.00	0.00E+00
	LOA-6	639.7	885.0	82.22	0.25	**2.56E+02**
	CSA-5	426.1	580.5	**0.62**	**0.01**	9.93E+02
	TruA-5	639.7	885.0	99.90	0.70	1.13E+03
	GCSA-6	483.7	664.3	0.74	0.06	4.02E+03
	ETAII-6	597.7	815.0	0.73	0.16	7.63E+03
	SCSA-6	316.8	436.1	0.72	0.16	7.64E+03
	ACAA-6	532.1	734.6	0.72	0.15	7.64E+03
	CCA-6	338.1	466.6	1.49	0.12	8.05E+03
	ESA-8	458.6	633.9	49.83	2.70	3.26E+04
	CSPA-6	303.7	416.6	9.13	1.30	6.46E+04
	ACA-8	340.8	463.3	0.68	1.20	1.39E+06
Unsigned multipliers	Accurate	2009.1	2784.7	0.00	0.00	0.00E+00
	TruM-5	1443.5	1997.0	99.89	4.47	**1.44E+12**
	AM2-14	1444.4	1998.4	99.10	**2.32**	4.02E+13
	TAM2-16	1459.2	2003.5	99.98	2.88	4.02E+13
	AM1-15	1412.1	1949.3	**98.81**	3.75	1.80E+14
	TAM1-16	1330.9	1818.9	99.99	6.45	2.02E+14
	PPAM-J0K13	1390.6	1917.7	99.99	245.96	3.20E+16
Signed multipliers	Accurate	2053.1	2827.3	0.00	0.00	0.00E+00
	TBM-7	1223.6	1684.4	99.99	39.23	3.86E+12

Table V shows the results for the circuits synthesized for low-power. Similarly, the automated designs are absent. Regarding the manual designs, LOA-6 shows the lowest MSE. However, CSA-5 has lower MRED and ER. Among the unsigned multipliers, TruM-5 shows the lowest MSE, while AM2-14 performs better in MRED. The TBM-7 is a good alternative for removing the temperature-induced guard-band of a signed multiplier.

Discussion: To determine the optimum solution, a large number of different approximate arithmetic circuits are evaluated. We demonstrated that different levels of approximation are required to overcome degradations under different degradation scenarios (aging or temperature) or circuit requirements (high-performance or low-power). Interestingly, we found that the truncation of LSBs is not always the most effective technique towards degradation-induced approximation. This is an important finding since current research work has been exclusively using this technique to trade-off degradations for a quality loss in the approximate arithmetic circuits.

We concluded for the high-performance approximate adders that CGPAs and LOA have the lowest error metrics among all the approximate circuits when we aimed to mitigate small degradations (aging-induced timing errors). Most of the approximate adders are designed for a high-speed operation (e.g., CSPA-5 and CSA-5), which made them suitable to overcome larger degradations (e.g., due to temperatures beyond 70°*C*). However, these designs generate higher power dissipation than using approximations in the LSBs [21]. The simulation results show similar trends for the approximate adders synthesized for low-power operation. Regarding the approximate multipliers, the truncation of LSBs has the lowest MSE towards degradation-induced approximation, independently of the workload scenario or circuit requirement. However, if the MRED is considered as the most important error metric instead of MSE, AM2, AM1 and TAM2 are more effective approximate designs.

(a) IDCT clk = 0.63 *ns* (b) Smoothing clk = 0.63 *ns* (c) Sharpening clk = 0.63 *ns*

Fig. 5. Results of three different image processing applications when the circuit is working in nominal conditions (25°*C*).

VI. ARCHITECTURE EXPLORATION RESULTS

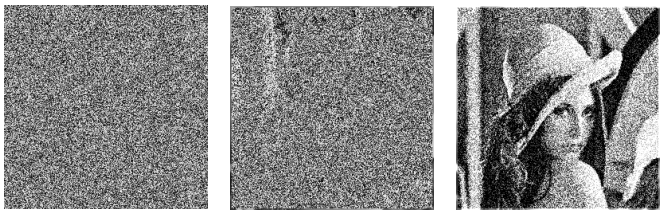
In this section, hardware accelerators for an inverse discrete cosine transform (IDCT), an image sharpening and an image smoothing application are implemented at the architecture level to improve the performance considering the effects of temperature. In the end, we demonstrated how to accurately exploit the degree of error tolerance for these applications, while still aiming to guarantee timing correctness without a performance loss.

A. Experimental Setup

In the scope of this evaluation, we aimed to mitigate guard-bands for a temperature of 70°*C* under two different design constraints. Signed multipliers are used for the IDCT application, and unsigned multipliers for the smoothing and sharpening applications. The RTL designs for these applications are synthesized with the 45-nm Nangate technology library [27] using the Synopsys Design Compiler. During the post-stress phase, we ran STA with PrimeTime to obtain the maximum delay in the circuit after inducing temperature with the degradation-aware cell libraries [24]. The PSNR metric is used to evaluate the output quality of 10 representative image files during the validation stage. In this article, we aimed at an output of at least 30 *dB*, which has been commonly used as an quality constraint for image processing applications [4]. Note that, as discussed previously, our methodology can be used with different application-dependent quality metrics and constraints. We use PSNR with a level of 30 *dB* as a case study here to demonstrate our approach. Finally, to compare this approach against state-of-the-art guard-band techniques, the three applications are synthesized with accurate circuits using the degradation-aware synthesis approach proposed in [17].

B. Image Processing Results

First, as a motivational study, we evaluate the image processing applications towards temperature-induced delay. Fig. 5 shows the ideal outputs of an IDCT, image smoothing and image sharpening applications when the chip is working at the nominal temperature (25°*C*). However, Fig. 6 shows how the



(a) IDCT clk = 0.63 *ns* (b) Smoothing clk = 0.63 *ns* (c) Sharpening clk = 0.63 *ns*

Fig. 6. Results of three different image processing applications when the circuit is exposed at 70°C without a technique to overcome the transistor degradations. Gate-level simulations are executed using PrimeTime and ModelSim to obtain the timing failures.

output considerably degrades when the circuits are exposed to a temperature of $70°C$ and no remedy is employed in the circuits to counteract the slow-down of the transistors. By this means, we show that guard-bands are required to ensure reliability even for error-tolerant applications.

As we discussed in Section II, most of the current guard-band techniques incur a significant performance loss to deal with these catastrophic events. Our approach, rather than using delay guard-bands to sustain reliability, employs approximations to compensate the temperature effects. The timing reports obtained from the Synopsys tools indicate that the multiplier constrains the critical path in the three image processing applications. Hence, the pre-characterized libraries in Section V-B can be employed to mitigate the temperature-induced delay with approximations during the design exploration stage (see methodology in Fig. 2). Table IV shows that truncation of 7 bits in the unsigned and signed multipliers is the best option to mitigate delay guard-bands at $70°C$. Following the proposed methodology, the accurate circuits are replaced by approximate circuits and then re-synthesized to optimize the surrounding glue logic. Note here that the timing constraint used during the synthesis process has to be modified to the smallest value of the new approximate circuit. Otherwise, the Design Compiler will relax the timing constraint and improve the other circuit measures rather than decreasing the critical path delay.

Table VI shows the circuit measures and output quality for the three applications. It should be noted that the delay column will determine the maximum clock frequency while still guarantying reliability in the circuit. However, a negative value in the slack column indicates that the design does not achieve the constrained timing. Therefore, a negative slack value can also be interpreted as the required delay guard-band in the circuit to avoid timing violations. Regarding the IDCT application, simulation results not only indicate that the TBM-7 meets the requirement of 30 *dB*, but also this approximate scheme completely remove delay guard-bands with positive gains in performance compared with the accurate circuit using the degradation-aware synthesis. Interestingly, we found that TruM-7 considerably degrades the output quality down to 6.76 *dB* and 18.56 *dB* for the smoothing and sharpening applications, respectively. Considering that performance is the most critical aspect for these architectures, we explored other approximate circuits prior to increasing the precision of TruM at the cost of small guard-bands. In this process, we found that the TAM1-16 [7] improves the output quality significantly while still meeting the delay requirement. Although these gains in output quality come at the expenses of an increase in power compared to TruM-7, we still observed overall positive gains (in delay, area and power) compared to the degradation-aware synthesis methodology [17].

In the second experiment, we repeated the process but then considered a different design constraint (low-power) for the same RTL applications. Although the delay guard-bands could be mitigated by using faster logic gates in low-power architectures, this approach leads to higher area and power consumption. In contrast, our approach maintains the performance without affecting other circuit metrics. Table VII shows the simulation results for the three image processing

TABLE VI
MEASURES FOR THE HIGH-PERFORMANCE APPLICATIONS RUNNING AT $70°C$

Application	**Delay** (*ns*)	**Slack**† (*ns*)	**Area** (mm^2)	**Power** (*mW*)	**PDP** (*pJ*)	**PSNR**†† (*dB*)
IDCT:						
Accurate*	0.86	-0.23	27.65	31.30	26.81	43.49
TBM-7‡	0.63	0.00	18.33	30.90	19.47	30.19
TBMS-8‡	0.78	-0.15	25.50	34.00	26.52	31.64
Smoothing:						
Accurate*	0.84	-0.21	3.64	2.98	2.50	Inf
TruM-7‡	0.63	0.00	1.98	2.14	1.34	6.76
TAM1-16‡	0.64	-0.01	2.41	2.36	1.50	37.27
Sharpening:						
Accurate*	0.87	-0.24	4.50	3.50	3.04	Inf
TruM-7‡	0.65	-0.02	2.40	2.62	1.71	18.56
TAM1-16‡	0.66	-0.03	2.85	2.92	1.91	45.56

† The required time is defined by the circuit without any degradations. †† Average of ten different images commonly found in multimedia applications. * The RTL implementation employs degradation-aware synthesis [17]. ‡ The RTL implementation employs our degradation-induced applications.

TABLE VII
MEASURES FOR THE LOW-POWER APPLICATIONS RUNNING AT $70°C$

Application	**Delay** (*ns*)	**Slack**† (*ns*)	**Area** (mm^2)	**Power** (*mW*)	**PDP** (*pJ*)	**PSNR**†† (*dB*)
IDCT:						
Accurate*	2.46	-0.52	24.68	10.90	26.78	43.49
TBM-7‡	1.58	0.36	16.69	11.60	18.38	30.19
TBMS-8‡	2.05	-0.11	24.00	13.00	26.61	31.64
Smoothing:						
Accurate*	2.50	-0.44	2.63	0.77	1.93	Inf
TruM-5‡	2.17	-0.11	1.60	0.56	1.21	17.10
TAM2-16‡	2.06	0.00	1.73	0.62	1.27	37.17
Sharpening:						
Accurate*	2.49	-0.40	3.58	1.01	2.51	Inf
TruM-5‡	2.13	-0.04	2.28	0.78	1.66	35.62
TAM2-16‡	2.07	0.02	2.28	0.75	1.55	46.44

† The required time is defined by the circuit without any degradations. †† Average of ten different images commonly found in multimedia applications. * The RTL implementation employs degradation-aware synthesis [17]. ‡ The RTL implementation employs our degradation-induced applications.

(a) PSNR = 48.69 *dB* clk = 0.86 *ns* (b) PSNR = 31.40 *dB* clk = 0.63 *ns* (c) PSNR = 32.24 *dB* clk = 0.78 *ns*

Fig. 7. IDCT outputs when the high-performance chip is exposed at 70°C using: (a) degradation-aware synthesis with an accurate circuit, (b) the approximate approach with TBM-7, and (c) the approximate approach with TBMS-8.

(a) PSNR = Inf *dB* clk = 0.84 *ns* (b) PSNR = 37.27 *dB* clk = 0.64 *ns* (c) PSNR = 6.94 *dB* clk = 0.63 *ns*

Fig. 8. Image smoothing outputs when the high-performance chip is exposed at 70°C using: (a) degradation-aware synthesis with an accurate circuit, (b) the approximate approach with TAM1-16, and (c) the approximate approach with TruM-7.

(a) PSNR = Inf *dB* clk = 0.87 *ns* (b) PSNR = 45.54 *dB* clk = 0.66 *ns* (c) PSNR = 17.56 *dB* clk = 0.65 *ns*

Fig. 9. Image sharpening outputs when the high-performance chip is exposed at 70°C using: (a) degradation-aware synthesis with an accurate circuit, (b) the approximate approach with TAM1-16, and (c) the approximate approach with TruM-7.

applications synthesized for low-power. Similar to the high-performance applications, the TBM-7 meets the requirement of 30 *dB* and the timing goal for the IDCT architecture. Regarding the sharpening and smoothing applications, TruM-5 design has the lowest MSE towards temperature-induced delay at 70°*C* for the unsigned multipliers (see Table V). However, the final output quality of the image processing applications is extremely low with this approximate scheme. Although the TAM2-16 and TruM-5 show similar characteristics in terms of circuit metrics at the architecture level, TAM2-16 generates a much better output quality especially in the smoothing application.

Figs. 7, 8 and 9 show a visual comparison of the output images for the IDCT, image smoothing and sharpening applications, respectively. As mentioned, the main objective for this methodology is to accurately trade-off delay guard-bands for a loss in output quality. Despite degradation-aware synthesis shows the best quality output employing accurate circuits, this approach incurs a penalty in hardware performance with the addition of delay guard-bands. On the other hand, the methodology not only mitigates delay guard-bands with a minimum reduction in the output quality, but also improves other circuit metrics such as area and power efficiency.

VII. Conclusions

Although guard-banding has been effectively applied to avoid degradation-induced timing errors, it directly impacts the power consumption or operating frequency of an integrated circuit. With the CMOS technology pushed to its physical limits to improve performance, improving reliability has become extraordinarily challenging. Therefore, in this article a complete framework is proposed to mitigate or completely remove guard-bands using the principles of approximate computing. Specifically, a methodology is developed to accurately convert degradations into controllable errors at the circuit level. A novelty in this article lies in the evaluation of a large number of approximate arithmetic circuits to determine the optimal solution and generate insights for mitigating the degradations due to aging and temperature effects.

The experiments show that a truncated adder is not the most effective technique, although it has been extensively used in the current literature. Among all the approximate adders, automatically generated adders using CGP produce the lowest error when we aimed to mitigate small degradations (aging-induced timing errors) followed by LOAs. Most of the approximate adders are designed for a high-speed by cutting the carry chain, which makes them suitable to overcome larger degradations such as high temperatures. Of the considered approximate multipliers, the truncated multiplier is the most effective design with respect to MSE. However, AM2, AM1 and TAM2 are the most effective approximate designs when the MRED is considered as the error metric.

Based on the pre-characterization of degradations at the circuit level, the degradations at the architecture level can effectively be dealt with. For three different image processing applications, the experiments reveal that temperature-induced degradation leads to an unacceptable quality loss, even for error-tolerant applications. Compared with the designs in literature, different approximation techniques are explored in more detail to trade-off guard-bands for approximations. The simulation results show that the MSE obtained at the circuit level is very relevant to application-specific metrics such as the PSNR. For the signed multipliers, we demonstrated that guard-bands are not only completely removed towards temperature-induced approximation, but also a gain of 28% in the PDP is achieved compared with the state-of-the-art approach from [17]. Similarly, we demonstrated that the TAM1 and TAM2 are the most effective schemes towards guard-band mitigation for the high-performance and low-power applications, respectively.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, vol. 38, no. 8, pp. 114–114, 1965.

[2] H. Amrouch and J. Henkel, "Containing guardbands," in *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan 2017, pp. 537–542.

[3] J. Han and M. Orshansky, "Approximate computing: an emerging paradigm for energy-efficient design," in *ETS*, 2013, pp. 1–6.

[4] H. Amrouch, B. Khaleghi, A. Gerstlauer, and J. Henkel, "Towards aging-induced approximations," in *ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2017, pp. 1–6.

[5] B. Boroujerdian, H. Amrouch, J. Henkel, and A. Gerstlauer, "Trading off temperature guardbands via adaptive approximations," in *2018 IEEE 36th International Conference on Computer Design (ICCD)*, Oct 2018, pp. 202–209.

[6] H. Kim, J. Kim, H. Amrouch, J. Henkel, A. Gerstlauer, K. Choi, and H. Park, "Aging compensation with dynamic computation approximation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 4, pp. 1319–1332, 2020.

[7] H. Jiang, C. Liu, F. Lombardi, and J. Han, "Low-power approximate unsigned multipliers with configurable error recovery," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 1, pp. 189–202, 2019.

[8] H. R. Mahdiani, A. Ahmadi, S. M. Fakhraie, and C. Lucas, "Bio-inspired imprecise computational blocks for efficient vlsi implementation of soft-computing applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 4, pp. 850–862, April 2010.

[9] H. Amrouch, "Techniques for aging, soft errors and temperature to increase the reliability of embedded on-chip systems," Ph.D. dissertation, CEA-LIST, Karlsruhe Institute of Technology, 2015.

[10] "Definition of: High-k/metal gate," accessed: 09 January 2020. [Online]. Available: http://www.pcmag.com/encyclopedia/term/58937/high-k-metal-gate

[11] M. Shafique, S. Garg, J. Henkel, and D. Marculescu, "The EDA challenges in the dark silicon era," in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2014, pp. 1–6.

[12] M. A. Scarpato, "Digital circuit performance estimation under PVT and aging effects," Ph.D. dissertation, Université Grenoble Alpes, 2017.

[13] M. Ebrahimi, F. Oboril, S. Kiamehr, and M. B. Tahoori, "Aging-aware logic synthesis," in *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2013, pp. 61–68.

[14] L. Zhang and R. P. Dick, "Scheduled voltage scaling for increasing lifetime in the presence of NBTI," in *2009 Asia and South Pacific Design Automation Conference*, Jan 2009, pp. 492–497.

[15] S. Das, C. Tokunaga, S. Pant, W. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. T. Blaauw, "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.

[16] V. M. van Santen, H. Amrouch, and J. Henkel, "New worst-case timing for standard cells under aging effects," *IEEE Transactions on Device and Materials Reliability*, vol. 19, no. 1, pp. 149–158, March 2019.

[17] H. Amrouch, B. Khaleghi, A. Gerstlauer, and J. Henkel, "Reliability-aware design to suppress aging," in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2016, pp. 1–6.

[18] H. Amrouch, S. Mishra, V. van Santen, S. Mahapatra, and J. Henkel, "Impact of BTI on dynamic and static power: From the physical to circuit level," in *2017 IEEE International Reliability Physics Symposium (IRPS)*, April 2017, pp. CR–3.1–CR–3.6.

[19] H. Amrouch, B. Khaleghi, and J. Henkel, "Optimizing temperature guardbands," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, March 2017, pp. 175–180.

[20] M. Sadi, G. K. Contreras, J. Chen, L. Winemberg, and M. Tehranipoor, "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, Nov 2017.

[21] H. Jiang, F. J. H. Santiago, H. Mo, L. Liu, and J. Han, "Approximate arithmetic circuits: A survey, characterization, and recent applications," *Proceedings of the IEEE*, vol. 108, no. 12, pp. 2108–2135, 2020.

[22] "Synopsys eda tools," accessed: 2019-06-28. [Online]. Available: http://www.synopsys.com/

[23] J. Bhasker and R. Chadha, "Static timing analysis for nanometer designs: A practical approach," *Springer*, New York, NY, 2009.

[24] "Degradation-Aware Cell Libraries, V1.0." accessed: 218-10-18. [Online]. Available: http://ces.itec.kit.edu/dependable-hardware.php

[25] J. Liang, J. Han, and F. Lombardi, "New metrics for the reliability of approximate and probabilistic adders," *IEEE Transactions on Computers*, vol. 62, no. 9, pp. 1760–1771, 2013.

[26] V. Chandra, "Monitoring reliability in embedded processors - a multi-layer view," in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2014, pp. 1–6.

[27] "Nangate, Open Cell Library." accessed: 218-10-25. [Online]. Available: http://www.nangate.com/

[28] H. Jiang, F. Hernandez, S. Ansari, L. Liu, B. Cockburn, F. Lombardi, and J. Han, "Characterizing approximate adders and multipliers optimized under different design constraints," *Great Lakes Symposium on VLSI*, pp. 393–398, 2019.

[29] W. Liu, J. Xu, D. Wang, C. Wang, P. Montuschi, and F. Lombardi, "Design and evaluation of approximate logarithmic multipliers for low power error-tolerant applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 9, pp. 2856–2868, 2018.

[30] M. S. Ansari, B. F. Cockburn, and J. Han, "An improved logarithmic multiplier for energy-efficient neural computing," *IEEE Transactions on Computers*, vol. 70, no. 4, pp. 614–625, 2021.

[31] W. Liu, L. Qian, C. Wang, H. Jiang, J. Han, and F. Lombardi, "Design of approximate radix-4 booth multipliers for error-tolerant computing," *IEEE Transactions on Computers*, vol. 66, no. 8, pp. 1435–1441, 2017.

[32] S. Venkatachalam, E. Adams, H. J. Lee, and S.-B. Ko, "Design and analysis of area and power efficient approximate booth multipliers," *IEEE Transactions on Computers*, vol. 68, no. 11, pp. 1697–1703, 2019.

[33] V. Mrazek, R. Hrbacek, Z. Vasicek, and L. Sekanina, "Evoapproxsb: Library of approximate adders and multipliers for circuit design and benchmarking of approximation methods," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, March 2017, pp. 258–261.

[34] Ning Zhu, W. L. Goh, and K. S. Yeo, "An enhanced low-power high-speed adder for error-tolerant application," in *Proceedings of the 2009 12th International Symposium on Integrated Circuits*, Dec 2009, pp. 69–72.

[35] K. Du, P. Varman, and K. Mohanram, "High performance reliable variable latency carry select addition," in *2012 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2012, pp. 1257–1262.

[36] D. Mohapatra, V. K. Chippa, A. Raghunathan, and K. Roy, "Design of voltage-scalable meta-functions for approximate computing," in *2011 Design, Automation Test in Europe*, March 2011, pp. 1–6.

[37] A. B. Kahng and S. Kang, "Accuracy-configurable adder for approximate arithmetic designs," in *DAC Design Automation Conference 2012*, June 2012, pp. 820–825.

[38] I. Lin, Y. Yang, and C. Lin, "High-performance low-power carry speculative addition with variable latency," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 9, pp. 1591–1603, Sep. 2015.

[39] Y. Kim, Y. Zhang, and P. Li, "Energy efficient approximate arithmetic for error resilient neuromorphic computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 11, pp. 2733–2737, Nov 2015.

[40] A. K. Verma, P. Brisk, and P. Ienne, "Variable latency speculative addition: A new paradigm for arithmetic circuit design," in *Proceedings of the conference on Design, automation and test in Europe*. ACM, 2008, pp. 1250–1255.

[41] J. Hu and W. Qian, "A new approximate adder with low relative error and correct sign calculation," in *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2015, pp. 1449–1454.

[42] L. Li and H. Zhou, "On error modeling and analysis of approximateadders,," *ICCAD*, pp. 511–518, 2014.

[43] G. Zervakis, K. Tsoumanis, S. Xydis, D. Soudris, and K. Pekmestzi, "Design-efficient approximate multiplication circuits through partial product perforation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 10, pp. 3105–3117, Oct 2016.

[44] V. Mrazek, Z. Vasicek, L. Sekanina, H. Jiang, and J. Han, "Scalable construction of approximate multipliers with formally guaranteed worst case error," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 11, pp. 2572–2576, Nov 2018.

[45] A. Momeni, J. Han, P. Montuschi, and F. Lombardi, "Design and analysis of approximate compressors for multiplication," *IEEE Transactions on Computers*, vol. 64, no. 4, pp. 984–994, April 2015.

[46] C. Lin and I. Lin, "High accuracy approximate multiplier with error correction," in *2013 IEEE 31st International Conference on Computer Design (ICCD)*, Oct 2013, pp. 33–38.

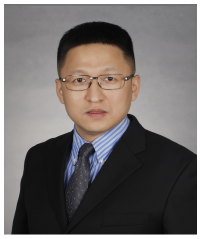
[47] P. Kulkarni, P. Gupta, and M. Ercegovac, "Trading accuracy for power with an underdesigned multiplier architecture," in *2011 24th Internatioal Conference on VLSI Design*, 2011, pp. 346–351.

Francisco Javier Hernandez Santiago received the B.Eng. degree in electronics and communications from the University of Guadalajara, Guadalajara, Mexico, in 2012, and the M.Sc. degree in computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2020. He is currently a Security Validation Engineer with Intel, Zapopan, Mexico. His research interests include approximate computing, machine learning, and computer security.

Honglan Jiang (Member, IEEE) received the B.Sc. and Master degrees in instrument science and technology from the Harbin Institute of Technology, Harbin, Heilongjiang, China, in 2011 and 2013, respectively. In 2018, she received the Ph.D. degree in integrated circuits and systems from the University of Alberta, Edmonton, AB, Canada. From 2018 to 2021, she worked as a postdoctoral fellow with the School of Integrated Circuits, Tsinghua University, Beijing, China. She is currently an associate professor with the Department of Micro-Nano Electronic, Shanghai Jiao Tong University, Shanghai, China. Her research interests include approximate computing, reconfigurable computing, and stochastic computing.

Hussam Amrouch (Member, IEEE) is a Junior Professor heading the Semiconductor Test and Reliability (STAR) chair within the Computer Science, Electrical Engineering Faculty at the University of Stuttgart as well as a Research Group Leader at the Karlsruhe Institute of Technology (KIT), Germany. He received his Ph.D. degree with highest distinction (Summa cum laude) from KIT in 2015. He servers currently as an Editor in the Nature Portfolio for the Nature Scientific Reports journal. His main research interests are design for reliability and testing from device physics to systems, machine learning, security, approximate computing, and emerging technologies with a special focus on ferroelectric devices. He holds eight HiPEAC Paper Awards and three best paper nominations at top EDA conferences: DAC'16, DAC'17 and DATE'17 for his work on reliability. He also serves as Associate Editor at Integration, the VLSI Journal. He has served in the technical program committees of many major EDA conferences such as DAC, ASP-DAC, ICCAD, etc. and as a reviewer in many top journals like Nature Electronics, TED, TCAS-I, TVLSI, TCAD, TC, etc. He has 170 publications in multidisciplinary research areas across the entire computing stack, starting from semiconductor physics to circuit design all the way up to computer-aided design and computer architecture. He is a member of the IEEE.

Andreas Gerstlauer (Senior Member, IEEE) received the Ph.D. degree in Information and Computer Science (ICS) from the University of California, Irvine (UCI) in 2004. Prior to joining UT Austin in 2008, he was an Assistant Researcher in the Center for Embedded Computer Systems (CECS) at UC Irvine. He is currently a Professor of Electrical and Computer Engineering (ECE) at The University of Texas at Austin. His research interests include system-level design automation, system modeling, design languages and methodologies, and embedded hardware and software synthesis. He is co-author on 3 books and more than 150 conference and journal publications. His work was recognized with several best paper awards and nominations, DAC, DATE and HOST, and as one of the most influential contributions in 10 years at DATE in 2008. He is the recipient of a 2016-2017 Humboldt Research Fellowship. He serves or has served as the General or Program Chair for major international conferences such as ESWEEK, MEMOCODE, CODES+ISSS and SAMOS; and as Associate and Guest Editor for ACM TECS and TODAES journals.



Leibo Liu (Senior Member, IEEE) received the B.S. degree in electronic engineering and the Ph.D. degree with the Institute of Microelectronics, both from Tsinghua University, Beijing, China, in 1999 and 2004, respectively. He is currently a Full Professor with the School of Integrated Circuits, Tsinghua University. His current research interests include reconfigurable computing, mobile computing, and very large-scale integration digital signal processing.

Jie Han (Senior Member, IEEE) received the B.Sc. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree from the Delft University of Technology, The Netherlands, in 2004. He is currently a Professor and Program Director of Computer Engineering in the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. His research interests include approximate computing, stochastic computing, reliability and fault tolerance, nanoelectronic circuits and systems, novel computational models for nanoscale and biological applications. He was a recipient of the Best Paper Award at the International Symposium on Nanoscale Architectures (NANOARCH 2015), and Best Paper Nominations at the 25th Great Lakes Symposium on VLSI (GLSVLSI 2015), NanoArch 2016, the 19th International Symposium on Quality Electronic Design (ISQED 2018) and the Design, Automation and Test in Europe Conference (DATE 2022). He was nominated for the 2006 Christiaan Huygens Prize of Science by the Royal Dutch Academy of Science. His work was recognized by *Science*, for developing a theory of fault-tolerant nanocircuits in 2005. He served as a General Chair for NANOARCH 2021, GLSVLSI 2017 and the IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT 2013); and a Technical Program Committee Chair for NANOARCH 2022, GLSVLSI 2016, DFT 2012, and the Symposium on Stochastic and Approximate Computing for Signal Processing and Machine Learning in 2017. He serves (or has served) as an Associate Editor for the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING (TETC), the IEEE TRANSACTIONS ON NANOTECHNOLOGY, the IEEE Circuits and Systems Magazine, the IEEE OPEN JOURNAL OF THE COMPUTER SOCIETY, Microelectronics Reliability and the Journal of Electronic Testing: Test and Application (JETTA, Elsevier).