Design, Verification, Test and In-Field Implications of Approximate Computing Systems

Alberto Bosio, Stefano Di Carlo, Patrick Girard, Ernesto Sanchez, Alessandro Savino, Lukas Sekanina, Marcello Traiola, Zdeněk Vašíček, Arnaud Virazel

Design, Verification, Test and In-Field Implications of Approximate Computing Systems

A. Bosio[2], S. Di Carlo[1], P. Girard[3], E. Sanchez[1], A. Savino[1], L. Sekanina[4], M. Traiola[2], Z. Vasicek[4], A. Virazel[3]
[1]*Control and Computer Eng. Dep., Politecnico di Torino*, Torino, Italy {alessandro.savino, ernesto.sanchez, stefano.dicarlo}@polito.it
[2]*École Centrale de Lyon*, Lyon, France {marcello.traiola,alberto.bosio}@ec-lyon.fr
[3]*LIRMM, Université de Montpellier / CNRS* , Montpellier, France {girard,virazel}@lirmm.fr
[4]*Faculty of Information Technology, Brno University of Technology*, Brno, Czechia {sekanina,vasicek}@fit.vutbr.cz

***Abstract*—Today, the concept of approximation in computing is becoming more and more a "hot topic" to investigate how computing systems can be more energy efficient, faster, and less complex. Intuitively, instead of performing exact computations and, consequently, requiring a high amount of resources, Approximate Computing aims at selectively relaxing the specifications, trading accuracy off for efficiency. While Approximate Computing gives several promises when looking at systems' performance, energy efficiency and complexity, it poses significant challenges regarding the design, the verification, the test and the in-field reliability of Approximate Computing systems. This tutorial paper covers these aspects leveraging the experience of the authors in the field to present state-of-the-art solutions to apply during the different development phases of an Approximate Computing system.**

***Index Terms*—approximate computing, circuit, design, test.**

I. Introduction

The reliance of the society on the use of information and communications technology (ICT) devices and systems is ever increasing. From the proliferation of e-mail and electronic document exchange, social media and apps to the ready use of mobile devices (already in their fourth generation), data analytics, and advanced computing to solve big challenges, ICT is having a disruptive impact on our society [1]. However, the energy consumption from the expanding use of ICT is unsustainable with present drivers, and it will impact heavily on the future climate change.

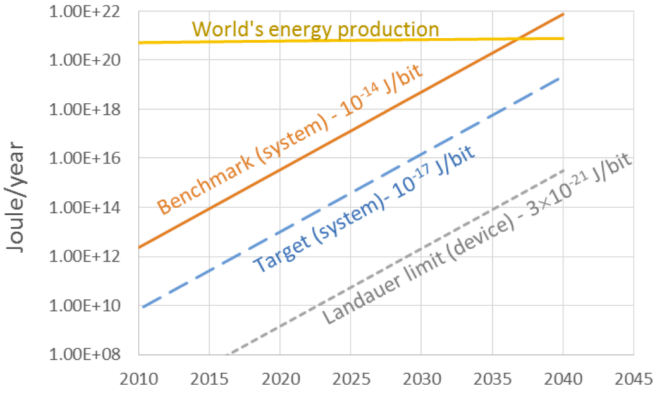


Fig. 1: Energy consumption trend in computing vs. the world energy production. Source: SIA/SRC [2]

Following the current trend, by 2040 computers will need more electricity than the world energy resources can generate, as illustrated in Figure 1. Already by 2025, data centers alone will consume 20% of all available electricity [3]. A similar trend exists on the communications side where, for example, energy consumption in mobile broadband networks and mobile terminals is comparable to data centers. In addition to the traditional personal communications, the Internet-of-Things (IoT) will soon connect up to 50 billion devices through wireless networks to the cloud, which will accelerate these trends [4].

Approximate and transprecision (i.e., adaptive precision) computing combined with application-specific processing structures are emerging computing paradigms able to support achieving the required energy efficiency improvements.

Since energy consumption (computing or communication) is the product of time and average power consumption of the device while carrying out an operation, these two factors, time and power, must be optimized for achieving energy savings. Approximate computing foresees to achieve this goal by considering a precious third design dimension, i.e., accuracy. The rationale at the base of this computing paradigm is that, in several parts of the global data acquisition, transfer, computation, and storage systems, it is possible to trade-off accuracy to either less power or less time consumed - or both.

As examples, numerous sensors are measuring noisy or inexact inputs; the algorithms processing the acquired signals can be stochastic; the applications using the data may be satisfied with an "acceptable" accuracy instead of exact and absolutely correct results; the system may be resilient against occasional errors; and a coarse classification or finding the most probable matches may be enough for a data mining system [5]–[7]. By introducing a new dimension – accuracy – to the design optimization, the energy efficiency can even be improved by a factor of 10x-50x.

While Approximate Computing gives several promises when looking at systems' performance, energy efficiency and complexity, it poses significant challenges regarding the design, verification, test and the in-field reliability of the approximated systems:

- *Design*: while several papers propose different approximation techniques at hardware and software level, the

decision of "what" to approximate and "how" to approximate given a target precision is still a challenging design space exploration problem that must be supported by dedicated design solutions and tools;

- *Verification and Testing*: verifying an approximated system and testing it at the end of production is a complex task. Traditional techniques for verification and testing start from the assumption that a system behaves in a deterministic way and any deviation from the planned behavior represents an hazard that must be addressed. This constraint is relaxed when Approximate Computing is applied, thus opening the path to different verification and testing techniques;
- *In-field*: once deployed in-field, approximated systems are still exposed to sources of errors (e.g., soft errors) like traditional precise systems. However, approximated systems have an intrinsic degree of error resilience, thus exposing inherent fault tolerance properties that can be exploited to reduce the reliability tax. This must be carefully budgeted when considering the reliability of the final system.

This paper overviews the above mentioned implications by presenting main challenges and state-of-art solutions derived from the application of Approximate Computing techniques in complex computing systems. The paper is structured following the main phases of the development cycle and use of a system: Section II considers the design phase, Section III the verification phase, Section IV the testing phase and Section V the implication of Approximate Computing on in-field operation. Finally, Section VI summarizes the main contributions of the paper.

II. Design phase

Several publications have contributed to the definition of different approximation techniques applicable at different design abstraction levels (e.g., gate level or architectural level) and different levels of the system stack (e.g., hardware level or software level). Every technique is in general characterized at the application level to understand the provided accuracy/implementation costs trade-offs. In general, the proposed approximation techniques can be grouped in three main categories [8]:

- *Software approximations*: approximate the software could correspond to functional approximations, such as a reduction in the number of iterations in an iterative-improvement algorithm [9], timing relaxation, and domain specific approximations. This approach has also proved to be resilient to fault, if needed [10].
- *Data approximations:* approximate data could be used to reduce the storage for data-intensive applications or for data resilient applications, such as neural networks and classifiers, to leverage the inherit error resiliency of the architecture [11], [12].
- *Hardware approximation*: approximate the hardware could replace or enhance the hardware layer with specific approximate components, such as adders and multipliers, to introduce intentional errors without modifying the algorithm [13]

When looking at evaluating a design, a primary concern is the fact that quality metrics may differ. Despite metrics such as delay/throughput, area, and power dissipation can be quantified at different levels of the design, and can still be comparable, the accuracy metric may be measured differently at various stages of design (refer to Figure 2). In fact, if for a multimedia application the Peak Signal-to-Noise Ratio (PSNR) or the Structural Similarity Index Metric (SSIM) can be a way to report the error, on other classification applications it can be the percentage of true positives [14]. If we move down in the system layers, at hardware level, the error metric can be evaluate as bit error rate (BER), or a probability distribution of error.

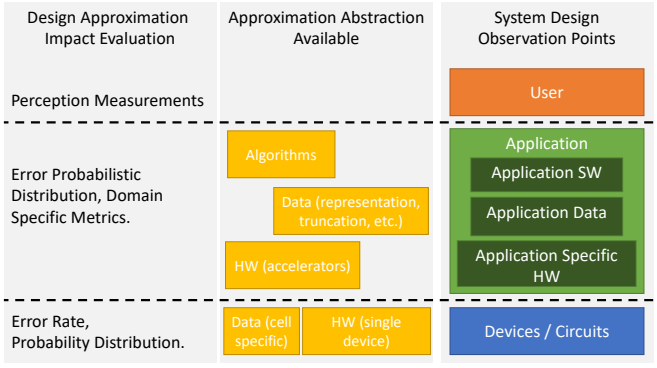


Fig. 2: Design approximation impact evaluation vs the levels of abstraction

It is clear that for the approximate computing to be effectively exploited within a design flow, it is essential to build quantitative approaches to model errors at each level of abstraction and to translate them to errors at other levels, in order to exploit the usage of the approximation at different levels of abstraction.

Another major concern for the approximate design regards the size of the design space to be explored. In fact, most of the research works already published address this problem by profiling the execution of the application several times, each time with a different version generated by combining one or more approximate operators together. The different design options are then compared with a golden approximate-free implementation [15]–[18]. This analysis can also be used to perform multi-objective optimization as proposed in [19] where authors exploit genetic programming for an automated functional approximation of combinational circuits at the gate and register-transfer levels.

In terms of multi-objective approaches, an interesting contribution is proposed in [20]. Authors reports a survey of approximation techniques with some relation with the energy efficiency of the computation, in a cross-layer fashion. Despite this work only addresses hardware components (adders and multipliers), it introduces the idea that the design should be able to consider different points of intervention within the full

stack of the system and also includes an experimental setup that addresses verification exploiting several versions.

A tool called IDEA, proposed in [21], moves the design to a higher level of abstraction by supporting a design space exploration based on the annotation of relaxing points into an application. Those relaxing points express the accuracy reduction constraints that are exploited by the tool to generate variants of a C/C++ application including different approximation techniques. Those variants are analyzed using a branch and bound technique. While this approach proposes a solution to the problem of automating the application of different approximate computing techniques to a program, it still resorts to the creation of variants and their execution to evaluate the impact of each approximation on the results based on benchmark specific metrics.

To cope with the timing require by the exploration of a huge design space, in [22] the authors propose a methodology for searching, selecting and combining the most suitable approximate circuits, from a set of available libraries, in order to generate an approximate accelerator. The methodology exploits machine learning techniques to generate several computational models of the accelerator. Each model is designed to ease the evaluation of the quality of the processing and the energy efficiency by means of a Pareto Frontier evaluated for each model. The whole approach is bounded to a given application but the machine learning approach reduces the time consuming exploration.

A broaden approach is described in [23], where the optimization starts from a RTL or HDL description of the hardware to optimize, thus still application-dependent, and its achieved by generating approximate high-level variants, through three steps: (i) transform the description into an Abstract Syntax Tree (AST) structure; (ii) create variants through transformations to the AST; and (iii) write the modified AST back into readable a RTL or HDL description. The interesting aspect of the approach is that the operators that can be targeted include data type simplifications, arithmetic operation approximations, arithmetic expressions transformations, variable-to-constant substitutions, and loop transformations. The final evaluation resorting to a multi-objective strategy inherited from NSGA-II [24].

Looking at all those contributions, the design phase seems to be burden more by an excess of alternatives to build and test than by the issue of properly evaluating them. For this reason, more sophisticated approaches, based on the stochastic properties of the approximation error have been proposed.

In [25], readers can find a first attempt to model the approximation resorting to statistic. The goal is to reduce the impact of the analysis by exploiting a circuit level model that makes feasible to characterize different approximate circuits. To effectively model the approximation in each circuit, the paper introduces the concept of *error profile*. The error profile resembles how, given the data distribution, the approximation error is introduced on the output of the approximate circuit.

Similar approaches have been proposed in [26]–[28], where specific implementations of hardware components, such as adders and multipliers are addressed. The basic idea is taking into account the different probability distributions of the input bits and evaluate the error distribution on the outcomes. All contributions point out that evaluating the error is faster than running several different versions of approximation with data patterns to evaluate the accuracy. Nevertheless, since all papers do not consider scenarios in which sequences of heterogeneous approximate operations are performed, a full exploitation of the stochastic approach is still not possible.

The propagation of the error is modeled in the approach described in [29], [30]. In the paper, authors report a formalization of the error introduced by different implementations of approximate operators and try to model its propagation within the application. The clear advantage is that the approach does not require several executions of the applications because the outcome reports the error distribution, with the only limitation of having the formalization application dependent.

In a more recent paper, [31], authors propose an error statistics evaluation for block-based approximate adders. The contribution is interesting because of the good characterization methodology for a single component that relies on a complete enumeration of all possible output deviations and the evaluation of their occurrences. Along with the limited types of approximate components addressed, the methodology does not consider the cumulative effects of the error propagation.

All previous probabilistic approaches are limited by the assumption regarding the data distribution of the inputs of the application. For more insight regarding what happens when the data distribution differs from the expected one, the reader can refer to [32]. Authors propose a study of the stability of the approximate circuits when the circuit targets a particular data distribution but the final workload differs from it.

This limitation, together with the proof that a probabilistic approach can simplify the evaluation of the accuracy of approximate designs, is what makes a stochastic approach able to support the design of approximate systems, allowing the assessment of the overall application error in a faster and reliable way. The idea is eventually addressed by few papers in the recent years [33], [34].

In order to model the effect of an approximate operation on the application result, these papers propose a stochastic approach based on a Bayesian Network (BN) model. The BN mimic the application with nodes representing data and operators, and arcs following the data-flow. The network makes it possible to model the error propagation along the data-flow of the application by populating the node with a set of probabilities of reporting the error distribution out of single approximate components. The accuracy assessment can be eventually done by estimating the error distribution of the application exploiting the Bayesian inference theory. Results are reported for several applications in which the approximation is obtained by scaling the precision of hardware operations and data registers. The main advantages of a fully stochastic methodologies are the need of profiling the application only once to construct the model, as well as, characterizing the operators only once and to be able to easily change the input

data distribution. Moreover, the methodology can effectively support the design exploration by giving an easy model to properly select the components of the application that might be worth approximate.

Finally, the design exploration can also take advantage of new metrics and strategies to select the components to be addressed by the approximation. Some very early and new research are exactly going in that direction [35]. The main idea is to anticipate the effect of the approximation by analyzing the data flow from the usage perspective, in order to further reduce the amount of different alternatives to be evaluated.

III. Verification phase

Determining the error of an approximate circuit or deciding whether an approximate circuit satisfies a given error constraint represent not only fundamental theoretical problems, but also highly practically relevant problems that must be routinely solved during the design of approximate circuits. While, at design time this task can be performed in an approximate way to support quick design space exploration, a precise analysis is required to verify that the final requirements of the system are met. A straightforward method to solve these problems is to use circuit simulation and *estimate* the error. If the *exact* error has to be determined then the circuit simulation has to be performed for all possible input vectors; however, this is applicable for small problem instances only. Hence, this section is focused on exact error analysis of approximate circuits by means of formal methods which are more scalable than circuit simulation in many cases. As arithmetic circuits frequently appear in the most popular error resilient applications (such deep learning and video processing), we focus on efficient exact error analysis of adders and multipliers. But the formal methods can be applied to effectively analyse errors of other combinational circuits (e.g., complex median networks [36]) as well as sequential systems [37].

Approximate implementations are usually created by (i) 'manual' modifications of exact circuits (see a detailed overview in [38]), (ii) developing new application-specific approximation schemes (see, e.g., new approximation techniques for FP multipliers [39]) or (iii) automated design space exploration algorithms [15], [19]. Fast and accurate error analysis is especially important in the case (iii) because the design space exploration methods sometime need to generate and evaluate millions of candidate design points.

A. Relaxed Equivalence Checking

Formal verification techniques that are widely adopted in the conventional circuit design flow are often based on *equivalence checking*, i.e., checking whether a mathematical model of a circuit under design meets a given specification. Two main approaches have been developed in this direction -– techniques based on Reduced Ordered Binary Decision Diagrams (ROBDD) and satisfiability (SAT) solvers [40]. In both cases, an auxiliary circuit, the so-called *miter*, is constructed and then analyzed. Fig. 3(a) shows that the miter instantiates both the candidate circuit F (to be checked) and

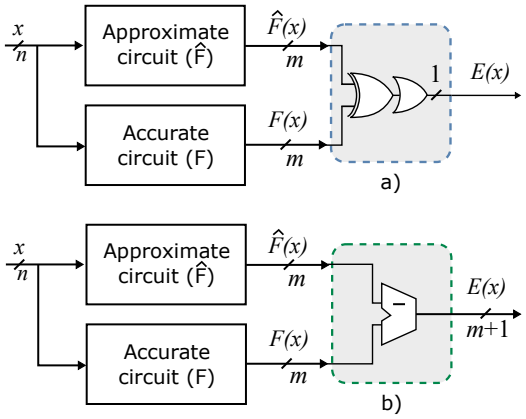


Fig. 3: Miter for equivalence checking (a) and arithmetic error analysis (b).

the golden circuit $\hat{F}$, and compares their corresponding outputs to detect a difference in their behaviour. In the context of approximate computing, we need to extend this concept to *relaxed equivalence checking*, by stressing the fact that the considered circuits will be checked to be equal up to some bound w.r.t. a suitably chosen distance (error) metric such as the worst case error and the average error. The (approximation) miter always contains an additional component enabling us to determine the error, see Fig. 3(b).

If the error analysis is performed using ROBDDs, a new ROBDD representing the miter is constructed by a procedure which reads the miter 'gate by gate' and adds appropriate nodes to ROBDD. ROBDDs can be directly used for the worst-case as well as the average-case analysis because every library for ROBDD manipulation is equipped with operations enabling us to address questions related to the satisfiability of the miter, namely finding one satisfying assignment and counting the number of satisfying assignments. The first operation provides a single input assignment x from the ON-set of a Boolean function. The second operation computes the size of the ON-set. As ROBDDs are inefficient in representing classes of circuits for which the number of nodes in BDD is growing exponentially with the number of input variables (e.g., multipliers and dividers), their use in relaxed equivalence checking is typically possible for adders and other less structurally complex functions. Anyway, for example, 128 bit adders can be quickly analysed in terms of all relevant error metrics [40].

If the error analysis is based on SAT solving, the miter is represented as a logic formula in Conjunctive Normal Form (CNF) for which SAT solver decides whether is satisfiable or unsatisfiable. The interpretation of this outcome depends on construction of the miter, see Section III-B. Common SAT solvers are, in principle, applicable to the worst-case analysis only. However, this approach is more scalable than ROBDDs for the error analysis of multipliers [19]. Specialized SAT solvers (#SAT) are capable of counting the number of satisfiable assignments, but their scalability is very limited

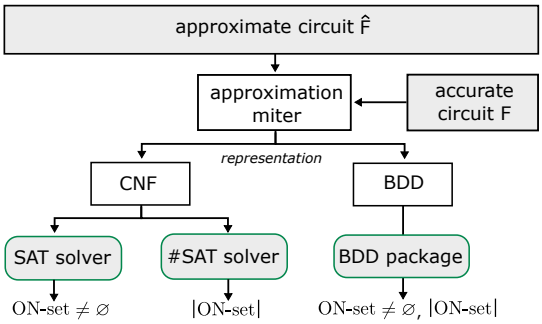


Fig. 4: Overview of formal error analysis approaches

and thus they are currently less practical for the exact error analysis [40].

B. Worst Case Error Analysis

The worst-case error analysis is typically based on an iterative approach in which a variant of binary search is applied.

Algorithm 1: Worst-case absolute error computation

Input: n-input approximation miter with m-bit signed output E in the two's complement
Output: maximum absolute arithmetic error (e_{wce})
$l \leftarrow 0$; $r \leftarrow 2^m - 1$
while $l \leq r$ **do**
  $t \leftarrow \lceil (l+r)/2 \rceil$
  if $\mathrm{WCEGT}(E, t)$ **then**
    $l \leftarrow t + 1$
  else
    $r \leftarrow t - 1$
return l

For computing the worst-case arithmetic error, for example, the miter given in Fig. 3(b) is used. Algorithm 1 illustrates the principle of determining the worst case arithmetic error, i.e. calculating the error magnitude at the m-bit output of the miter denoted as E. The principle of this procedure is to iteratively check whether the error is greater than a given threshold (denoted as t in the algorithm). The search procedure gradually narrows down the interval where the exact error value lies. After a finite number of steps, a single value is determined. As the binary search runs in logarithmic time with respect to the range, at most m comparisons are required. The checking can be ensured by means of the magnitude comparator which is used to form a Boolean function whose output is equal to 1 if and only if a given worst-case error $\mathcal{T}$ is violated by the circuit under analysis.

$$\begin{aligned} &\mathrm{WCEGT}(E, \mathcal{T}) = \exists_{x \in \mathbb{B}^n} |E(x)| > \mathcal{T} \\ &= \text{ON-set}\left([\overline{e_m} \wedge (E > \mathcal{T})] \vee [e_m \wedge (\overline{E} > (\mathcal{T} - 1))] \right) \neq \varnothing. \end{aligned} \tag{1}$$

Then, the satisfiability of this function can be investigated. An incremental SAT solver should be employed to mitigate a potential overhead caused by the necessity of constructing a different comparator in each iteration [40].

C. Average-case error analysis

Determining the average-case error represents a substantially harder problem because it requires the counting of the number of satisfiable assignments. For computing the average-case arithmetic error, for example, the same miter as in the previous case is used. The mean absolute error can be obtained by determining the error probability per each output bit. The obtained counts are then weighted according to the significance of the output bits and summed up. This is illustrated in Algorithm 2.

Algorithm 2: Mean absolute error computation

Input: n-input approximation miter with m-bit signed output e in the two's complement, i.e. $E = 2^m e_m - \sum_{i=0}^{m-1} 2^i e_i$
Output: mean absolute arithmetic error (e_{mae})
$\varepsilon, c \leftarrow |\,\text{ON-set}(e_m)|$
for $i \in \{0, 1, \ldots, m-1\}$ **do**
  if $c > 0$ **then**
    $\varepsilon \leftarrow \varepsilon + 2^i |\,\text{ON-set}(e_i \oplus e_m)|$
  else
    $\varepsilon \leftarrow \varepsilon + 2^i |\,\text{ON-set}(e_i)|$
return $2^{-n}\varepsilon$;

D. Comparison

Detailed analysis of relaxed equivalence checking algorithms has recently been performed in [40]. The analysis revealed that the computational complexity of the SAT-based methods heavily depends on the actual worst-case error. The computational time increases with a decreasing error, which is noticeable especially on multipliers. For example, tens of milliseconds are needed to analyze the 12-bit multipliers having the error higher than 2.7%. On the other hand, higher tens of seconds are needed for instances having the error in the range (0.37%, 2.71%] and no result was obtained for multipliers having the worst-case error below 0.05%.

Figure 5 shows the computational requirements of the WCEGT procedure (i.e. worst-case error checking) for five different thresholds applied to 8-bit multipliers. The worst-case error checking is extremely fast (few milliseconds are required) but only if the actual worst-case error (denoted as wce) is higher than a given threshold $\mathcal{T}$. If this condition is violated, the CPU time may increase by several orders of magnitude. Surprisingly, the difference between the worst case and the best case CPU time increases with decreasing the threshold $\mathcal{T}$. Performing WCEGT for thresholds below 1.5% represents the most difficult case. Up to 100 seconds are required to analyse the circuit instances whose wce is lower than the chosen threshold. Considering this fact, the design of multiplier-based approximate circuits with low error will be a challenging task because the checking will represent the bottleneck of the whole design process.

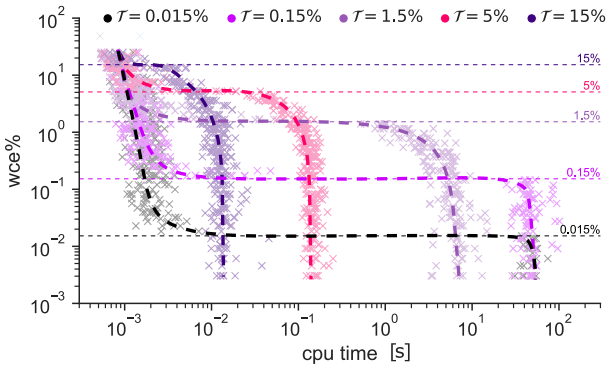


Fig. 5: The computational requirements of the WCEGT procedure proving that $e_{wce} > \mathcal{T}$ of 8-bit approximate multipliers taken from EvoApprox library.

IV. Testing phase

The application of approximate computing at hardware level results in systems widely referred to as *Approximate Integrated Circuits (AxICs)*. An extensively used method to design those circuits is functional approximation of conventional integrated circuits [41]. This section focuses specifically on the testing aspects of functionally approximate circuits. Indeed, since approximation changes the functional behavior of circuits, techniques to test them must be revisited. As a matter of fact, extending the basic testing concepts to AxICs is not straightforward. In particular, during the test of a conventional circuit, any change in its functional output signals with respect to the expected values leads to labeling the circuit as faulty, and discarding it. When moving to AxICs, the presence of a fault may lead the circuit to behave differently than expected, yet still in an acceptable manner. In this case the circuit should not be discarded. Mastering these mechanisms may lead to increase the production process yield.

This section presents a whole new test flow – called *Approximation-Aware (AxA) test flow* – to deal with such aspects. It is the result of several contributions in the last years [42]–[51]. The flow is composed of three main steps: (i) AxA fault classification, (ii) AxA test pattern generation and (iii) AxA test set application. Briefly, the *fault classification* divides faults producing catastrophic effects on the circuit behavior from those producing acceptable effects. The *test pattern generation* produces test stimuli able to cover all the catastrophic faults and, at the same time, to leave acceptable faults undetected, as much as possible. Finally, the *test set application* labels AxICs under test as catastrophically faulty, acceptably faulty, or fault-free. Only AxICs falling into the first group will be discarded, thus minimizing overtesting (i.e., minimizing AxICs discarded due to acceptable faults). Next subsections describe each AxA test step.

A. AxA fault classification

The first step of the AxA testing is the *fault classification*. It aims at separating acceptable faults from catastrophic ones. Moreover, fault classification establishes the *expected yield increase* of the AxA testing w.r.t. conventional test. Measuring the output deviations of AxICs is a crucial task for a successful classification. Different error metrics have been proposed in the literature to measure AxIC output deviations [52]. In [49], we showed that the classification task complexity drastically changes depending on the considered error metric. We showed that some metrics – referred to as *Single Condition Test (SCT) metrics* – entail a smaller effort for the fault classification compared to metrics based on the calculation of a mean – referred to as *Mean Error (ME) metrics*.

In [46], [49] we presented two fault classification techniques to address respectively SCT and ME metrics. Both techniques are based on the idea of masking acceptable fault effects by using a filter. Specifically, both the netlists of the AxIC under test and of the original precise circuit are embedded in a *classifying architecture*, along with the filter. For a given fault, the so-obtained architecture produces an anomaly only if the fault leads to catastrophic output deviations. In this way, by using conventional test approaches, it is finally possible to distinguish catastrophic faults from acceptable ones. The classifying architecture is never manufactured. It is only used in simulation to classify faults. Furthermore, the technique proposed in [46] entailed drastically reduced times compared to other state-of-the-art techniques [53], [54].

B. AxA test pattern generation

The second step of the AxA testing is the *test pattern generation*. In the context of AxICs, test patterns must cover all catastrophic faults and as few as possible acceptable ones. Respecting both these conditions is crucial to discard AxICs affected by catastrophic defects and, at the same time, to avoid discarding those affected by acceptable defects. Since state-of-the-art techniques [53], [54] do not focus on minimizing detected acceptable faults, in [50] we presented the first technique to suitably address the AxA test pattern generation.

This novel technique relies on a new engine capable of finding, among a set of input vectors, the smallest subset covering all the catastrophic faults and minimizing the acceptable fault coverage. Specifically, the engine generates an input vector set $\mathcal{S}$ and measures its catastrophic fault coverage as well as its acceptable fault coverage. Hence, it finds within $\mathcal{S}$ the optimal subset $\mathcal{V}$ which attains the required goals. To accomplish this task, the engine formulates and resolves an *Integer Linear Programming (ILP)* optimization problem, whose solution is the final *ax-aware test set*.

Experimental outcomes achieved with the proposed technique showed an improvement spanning from 16% to 49% compared to state-of-the-art techniques. Although the achieved results are quite good, the ideal outcomes (i.e., 100% covered catastrophic faults and 0% covered acceptable faults) were still quite far from being attained. Therefore, we dedicated further efforts to effectively test AxICs, as shown in next subsection.

C. AxA test set application

To push further the test outcomes, the third step of AxA testing, the *test pattern application*, comes into play. In this

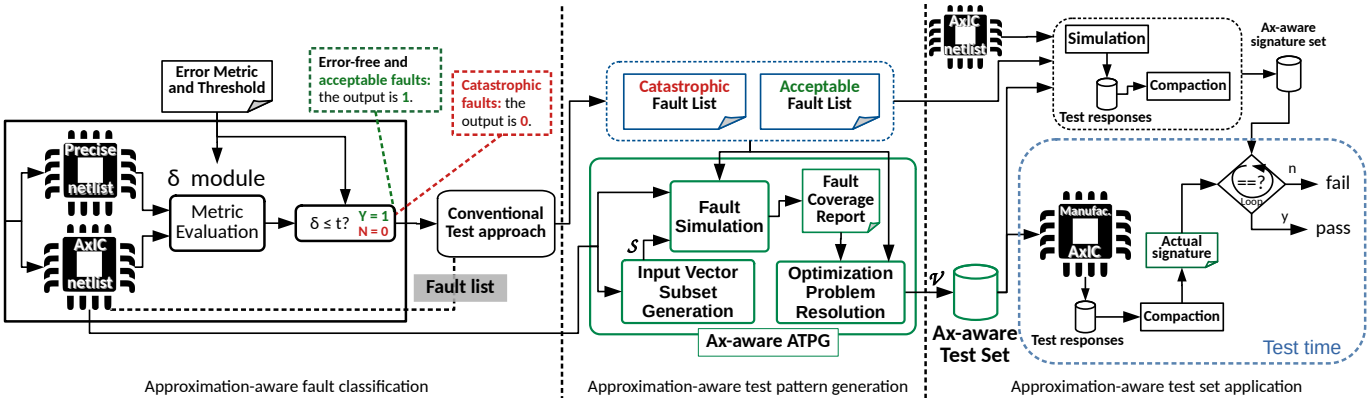


Fig. 6: Approximation-Aware (AxA) test flow

regard, in [51] we presented the AxA test set application technique. Since often it is not possible to avoid detecting acceptable faults, the basic idea is to verify, after the test application, whether the detected fault was acceptable or not.

The proposed technique is based on the well-know *signature analysis* concept, successfully applied to built-in self-test (BIST) architectures in the seventies [55] and still used in modern BIST architectures. The conventional signature analysis approach compacts test responses of a fault-free circuit into a *golden signature* (i.e., the reference behavior). In the test phase, the test responses of the circuit under test are compacted together into a signature (i.e., the actual behavior). Hence, the latter is compared with the golden one. If the two signatures are identical, the circuit under test is considered fault-free; otherwise, a malfunction is detected.

We drew inspiration from the signature analysis and proposed a technique divided into two steps:

At design time, we perform a fault simulation by using test patterns and the AxIC's faults. For each fault, we compact simulation responses into a signature. We obtain acceptable and catastrophic signatures. We remove from acceptable signatures those overlapping with catastrophic ones, thus ending up having an *ax-aware signature set*.

At test time, manufactured AxIC test responses are compacted into a signature and compared with the ones in the ax-aware signature set. If there is at least one match, then the AxIC is considered acceptable. Otherwise, the circuit is rejected.

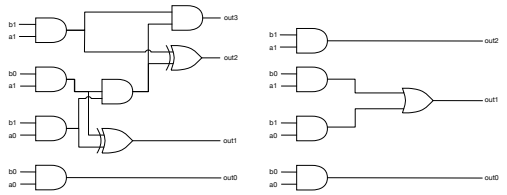
The proposed technique is intended to be used for external test, i.e., test are applied by using an Automatic Test Equipment (ATE). Of course, it can be also adapted to a BIST context.

Results obtained with the proposed technique were really good. Indeed, they showed yield gain results very close to the expected ones (i.e., 99.84% of the expectations, on average). In terms of covered faults, the technique delivered 100% covered catastrophic faults and 0.16% covered acceptable faults on average, that are very close to the ideal ones (i.e., 100% covered catastrophic faults and 0% covered acceptable faults).

V. In-field

As described before, Approximate Computing techniques have been positively introduced thanks to the intrinsic resilience of many applications [56]; as a collateral resiliency effect, it could be also stated that a resilient application has the capability to produce acceptable outputs despite underlying computations being affected by hardware faults.

As initially presented in [57], here we describe how Approximate Computing can positively impact the intrinsic circuits resilience by allowing faulty devices to work as approximate ones.



(a) Precise (b) Approximate

Fig. 7: Two-bit multiplier Example

Let us consider the accurate and approximate implementations of a 2-bit multiplier shown in Figure 7. Assume also to analyze the case when the given circuits are affected by a Stuck-at Fault (SaF). The presence of a fault f can lead to erroneous outputs characterized by a given WCE_f. If $WCE_f \leq WCE_{tr}$ the application can tolerate the presence of f, otherwise the application cannot tolerate f anymore.

We can classify the fault universe F_u in two subsets:

1) F_t: $\vee f_{ti} \in F_t \rightarrow WCE_{f_{ti}} \leq WCE_{tr}$
2) F_c: $\vee f_{ci} \in F_c \rightarrow WCE_{f_{ci}} > WCE_{tr}$

Where F_t is the set of *Tolerable* Faults, F_c is the set of *Critical* Faults. In [46], authors presented a methodology for identifying the two sets. Results are gathered in Table I.

TABLE I: Tolerable and Critical Faults

Circuit	#F_t	#F_c	#TV
Precise	25	23	4
AxC	23	7	3

It is possible to see in Table I that for the precise multiplier (Fig. 7a) only 23 SaFs are included in the Critical Faults set F_c. In other words, about half of the faults leads to an error lower than $WCE_{tr} = 2$, and in the case one of these errors appear, it is possible to ensure that the circuit will work as an approximate one.

Considering the approximate multiplier, it is possible to see that only 7 faults belong to the critical set. However, it is important to highlight that even though tolerable faults may do not impact the WCE, they can impact other metrics, for example, the Bit Error Rare (BER). This has to be carefully evaluated at application-level.

In order to analyze the possibility of using a faulty circuit as an approximate one, we follow a similar approach than the one presented in [46]. The main difference is that in this work we aim at identifying the set of tolerable faults w.r.t. to a given WCE. In our framework, the faulty circuit can be either a precise or an approximate one.

The whole process is composed of two steps: The first step is performed offline and consists of a preliminary analysis of the precise/approximate circuit in the presence of faults, it is necessary to define the acceptable WCE. During this step, the circuit Critical Faults (the ones that need to be mandatory tested), and Tolerable Faults (the ones that produce an acceptable lost in accuracy) are identified. Additionally, a test set is created able to test the critical faults.

The second step is performed in-field and requires to use a set of test patterns covering the set of Critical Faults. In the case the circuit is affected by one of this faults, it is not possible to accept the results since the WCE is higher than the permitted one. On the contrary, the circuit degradation is lower than the maximum admissible one, and the results can be considered as if produced by an approximate one.

We analyzed 7 widely used 8-bit precise adders synthesized using a 65 nm technology library: Carry Select ($CarrySel$), Ripple Carry ($RippCarry$), Carry Lookahead ($CarryLKH$), Higher Valency Tree Adder with HanCarlson Architecture ($HVTrHCA$), Higher Valency Tree Adder with Kogge-Stone Architecture ($HVTrKSA$), Tree Adder with HanCarlson Architecture ($TwHCA$), and Tree Adder with Kogge-Stone Architecture ($TwKSA$). Moreover, we also considered a set of approximate adders. Two techniques have been used for approximation: (i) Precision Reduction and, (ii) Functional Approximation. Precision Reduction approximate adders are obtained from the precise adders by simply cutting the four LSBs of each operand. For each precise adder, the approximate versions are obtained, and have the same number of faults as the precise ones since they have the same circuit netlist:

- 4 LSBs set to '0': leading to a WCE = 15.

The considered adders are available in [58]. We selected as examples on functional approximation, the approximate adders having WCE = 15 in order to be comparable with the approximation induced by the precision reduction made on the accurate adders.

Table II summarizes the obtained results when WCE = 15. For the precise adders (listed in the first column), we report the percentage of Critical Faults (CF) and Test Length (TL) (i.e., number of test vectors required to only detect the Critical Faults) depending on the adders' version: the $Precise$ implementation, the precision reduction approximation when truncated the 4 LSB ($TR4$). In addition, the very last columns of the table report also the results on the functional approximate adders (the approximate ones obtained from [58]) having WCE = 15.

(a) PR - TR4: SSIM = 0.904982

(b) FA - Add23: SSIM = 0.996768

Fig. 8: Approximate Adders Application Accuracy.

From the reported data it is observed that:

i) The amount of Critical Faults decreases from the precise version down to the approximate ones. Interestingly, certain functional approximate adders show a significantly lower percentage of Critical Faults, and as expected, also the test length is reduced.

ii) The approximation technique can definitely help to reduce the test time. The exploited approximation technique can lead to different results, precision reduction versus functional approximation. However, even for functional approximation, different 'kind' of approximation results in different test time. Since the exploited circuits are generated from an evolutionary algorithm, automated generation should be investigated with the goal of improving the percentage of critical faults and check (i.e., take as objective the generation of approximate circuits more 'resilient' than usual).

Results presented so far have been extracted by considering the adders as a stand-alone circuit. However, to really prove that faulty circuits (either precise or approximate) can be used as approximate ones, one application will be used as a case study and detailed in the next subsection.

A. NOVA

Nova is a low-power real-time H.264 Advanced Video Coding, targeting mobile applications. Nova HDL model is available from [59]. We replaced the adders of NOVA by the precise and approximate ones reported before. For each NOVA implementation, we perform a simulation-based fault injection. Fault injection consists in injecting one SaF at a time and check the impact of the fault at the application outputs. The injected faults are the critical and tolerable ones identified in the previous section. Workloads came from [60].

First of all, we check the impact of the approximate adders on the application accuracy. The latter is quantified by using the Structural SIMilarity (SSIM) index [61]. Since the NOVA output is a video, we compute the SSIM for each frame and

TABLE II: Critical Fault and Test Length when WCE = 15

Circuit	Version				Add_012		Add_013		Add_016		Add_023		Add_025		Add_40	
	Precise		TR4													
	CF	TL	CF	TL	CF	TL	CF	TL	CF	TL	CF	TL	CF	TL	CF	TL
CarryLKH	82.86%	41	65.31%	28												
CarrySel	78.05%	16	78.05%	15												
HVTrHCA	72.09%	19	61.24%	14												
HVTrKSA	86.44%	24	67.51%	19	90.29%	18	50.00%	4	80.77%	6	50.00%	4	99.21%	13	45.65%	5
RippCarry	53.85%	7	53.85%	7												
TwHCA	71.74%	19	65.22%	18												
TwKSA	85.45%	27	67.58%	17												

Fig. 9: Mispredicted Faults.

then we compute the average. Figure 8 depicts an example of videos obtained by using the Precision Reduction (PR) with 4 bits truncated. The figure presents frames of the videos obtained by using Functional Approximation (FA) Add_23 (WCE = 15). By simply looking at Fig. 8, the **PR - TR4** shown the worst quality. We can quantify the videos quality w.r.t. to the precise output by the SSIM as reported in the figures caption.

We also check the results when injecting the Critical and Tolerable faults identified previously. For example, Fig. 9 presents a screenshot of a video obtained by injecting a 'tolerable' fault in the Add_23, leading to a SSIM of 0.95. As it can be noticed, the resulting output is of unacceptable quality, this means that some faults classified as tolerable lead to a too low SSIM.

This result can be explained by the fact that the injected fault is classified as tolerable because it leads to a WCE lower than 15, but it causes other errors. For example, it increases the bit error rate. This clearly indicates that considering only one metric may be not enough, and the most important, the impact of a fault has to be analyzed considering the application.

On the other hand, the precise adders behave much better than the approximate ones, then these adders can be used as an excellent approximate one when tolerable faults appear. This clearly demonstrates that approximation can be really used to improve the system lifetime.

VI. Conclusions

This tutorial paper presented an overview of different approaches to handle the design, verification, testing and in-field operation of approximate computing systems. The presented solutions are not exhaustive and new publications and approaches will appear while the field becomes mature. The paper leverages on the experience of the authors to overview the major challenges that still represent a barrier to transform this interesting research field into real solutions ready to the market.

Acknowledgements

This work was supported by Czech Science Foundation project 19-10137S.

References

[1] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM journal of research and development*, vol. 5, no. 3, pp. 183–191, 1961.

[2] Semiconductor Industry Association and others, "Rebooting the it revolution: A call to action," [Online] https://www.src.org/newsroom/rebooting-the-it-revolution.pdf, 2015.

[3] J. Marques Lima, "Data centres of the world will consume 1/5 of earth's power by 2025," *Data Economy*, 2017. [Online]. Available: https://economy.com/data-centres-world-will-consume-1-5-earths-power-2025/

[4] Juniper Research, "Iot connections to grow 140% to hit 50 billion by 2022, as edge computing accelerates roi," [Online] https://www.juniperresearch.com/press/press-releases/iot-connections-to-grow-140-to-hit-50-billion, 2018.

[5] A. Sampson, A. Baixo, B. Ransford, T. Moreau, J. Yip, L. Ceze, and M. Oskin, "Accept: A programmer-guided compiler framework for practical approximate computing," *University of Washington Technical Report UW-CSE-15-01*, vol. 1, no. 2, 2015.

[6] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," in *Test Symposium (ETS), 2013 18th IEEE European*. IEEE, 2013, pp. 1–6.

[7] X. Wu, X. Zhu, G. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, Jan 2014.

[8] S. Mittal, "A survey of techniques for approximate computing," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 62:1–62:33, Mar. 2016.

[9] V. K. Chippa, S. Venkataramani, S. T. Chakradhar, K. Roy, and A. Raghunathan, "Approximate computing: An integrated hardware approach," in *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2013, pp. 111–117.

[10] G. S. Rodrigues, F. L. Kastensmidt, V. Pouget, and A. Bosio, "Performances vs reliability: how to exploit approximate computing for safety-critical applications," in *2018 IEEE 24th Int. Symposium on On-Line Testing And Robust System Design*, July 2018, pp. 291–294.

[11] A. Ranjan, S. Venkataramani, X. Fong, K. Roy, and A. Raghunathan, "Approximate storage for energy efficient spintronic memories," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2015, pp. 1–6.

[12] B. Barrois and O. Sentieys, "Customizing fixed-point and floating-point arithmetic — a case study in k-means clustering," in *2017 IEEE Int. Workshop on Signal Processing Systems*, Oct 2017, pp. 1–6.

[13] M. Macedo, L. Soares, B. Silveira, C. M. Diniz, and E. A. C. da Costa, "Exploring the use of parallel prefix adder topologies into approximate adder circuits," in *2017 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Dec 2017, pp. 298–301.

[14] A. G. M. Strollo and D. Esposito, "Approximate computing in the nanoscale era," in *2018 International Conference on IC Design Technology (ICICDT)*, 2018, pp. 21–24.

[15] S. Venkataramani, A. Sabne, V. J. Kozhikkottu, K. Roy, and A. Raghunathan, "SALSA: systematic logic synthesis of approximate circuits," in *The 49th Design Automation Conference*. ACM, 2012, pp. 796–801.

[16] S. Lee, L. K. John, and A. Gerstlauer, "High-level synthesis of approximate hardware under joint precision and voltage scaling," in *Design, Automation Test in Europe*, 2017, pp. 187–192.

[17] A. Sampson, W. Dietl, E. Fortuna, D. Gnanapragasam, L. Ceze, and D. Grossman, "Enerj: Approximate data types for safe and general low-power computation," *ACM SIGPLAN Notices*, vol. 46, no. 6, pp. 164–174, 2011.

[18] C. Rubio-González, C. Nguyen, H. D. Nguyen, J. Demmel, W. Kahan, K. Sen, D. H. Bailey, C. Iancu, and D. Hough, "Precimonious: Tuning assistant for floating-point precision," in *Proc. of the Int. Conf. on High Performance Computing, Networking, Storage and Analysis*. ACM, 2013, p. 27.

[19] L. Sekanina, Z. Vasicek, and V. Mrazek, *Automated Search-Based Functional Approximation for Digital Circuits*. Springer International Publishing, 2019, pp. 175–203.

[20] M. Shafique, R. Hafiz, S. Rehman, W. El-Harouni, and J. Henkel, "Invited: Cross-layer approximate computing: From logic to architectures," in *53nd Design Automation Conference*, 2016, pp. 1–6.

[21] M. Barbareschi, F. Iannucci, and A. Mazzeo, "A pruning technique for b b based design exploration of approximate computing variants," in *IEEE Computer Society Annual Symposium on VLSI*, 2016, pp. 707–712.

[22] V. Mrazek, M. A. Hanif, Z. Vasicek, L. Sekanina, and M. Shafique, "autoax: An automatic design space exploration and circuit building methodology utilizing libraries of approximate components," in *2019 56th ACM/IEEE Design Automation Conference (DAC)*, 2019, pp. 1–6.

[23] K. Nepal, S. Hashemi, H. Tann, R. I. Bahar, and S. Reda, "Automated high-level generation of low-power approximate computing circuits," *IEEE Transactions on Emerging Topics in Computing*, vol. 7, no. 1, pp. 18–30, 2019.

[24] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[25] R. Venkatesan, A. Agarwal, K. Roy, and A. Raghunathan, "Macaco: Modeling and analysis of circuits for approximate computing," in *2011 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2011, pp. 667–673.

[26] M. K. Ayub, O. Hasan, and M. Shafique, "Statistical error analysis for low power approximate adders," in *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2017, pp. 1–6.

[27] S. Mazahir, O. Hasan, R. Hafiz, and M. Shafique, "Probabilistic error analysis of approximate recursive multipliers," *IEEE Transactions on Computers*, vol. 66, no. 11, pp. 1982–1990, Nov 2017.

[28] S. Mazahir, O. Hasan, R. Hafiz, M. Shafique, and J. Henkel, "Probabilistic error modeling for approximate adders," *IEEE Transactions on Computers*, vol. 66, no. 3, pp. 515–530, March 2017.

[29] K. N. Parashar, D. Menard, and O. Sentieys, "Accelerated performance evaluation of fixed-point systems with un-smooth operations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 33, no. 4, pp. 599–612, April 2014.

[30] R. Rocher, D. Menard, P. Scalart, and O. Sentieys, "Analytical approach for numerical accuracy estimation of fixed-point systems based on smooth operations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 10, pp. 2326–2339, Oct 2012.

[31] Y. Wu, Y. Li, X. Ge, Y. Gao, and W. Qian, "An efficient method for calculating the error statistics of block-based approximate adders," *IEEE Transactions on Computers*, vol. 68, no. 1, pp. 21–38, Jan 2019.

[32] S. Xu and B. C. Schafer, "Exposing approximate computing optimizations at different levels: From behavioral to gate-level," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 11, pp. 3077–3088, 2017.

[33] M. Traiola, A. Savino, M. Barbareschi, S. D. Carlo, and A. Bosio, "Predicting the impact of functional approximation: from component- to application-level," in *24th Int. Symposium on On-Line Testing And Robust System Design*, July 2018, pp. 61–64.

[34] M. Traiola, A. Savino, and S. D. Carlo], "Probabilistic estimation of the application-level impact of precision scaling in approximate computing applications," *Microelectronics Reliability*, vol. 102, p. 113309, 2019.

[35] A. Savino, M. Portolan, R. Leveugle, and S. Di Carlo, "Approximate computing design exploration through data lifetime metrics," in *2019 IEEE European Test Symposium (ETS)*, 2019, pp. 1–7.

[36] Z. Vasicek and V. Mrazek, "Trading between quality and non-functional properties of median filter in embedded systems," *Genetic Programming and Evolvable Machines*, vol. 18, no. 1, pp. 45–82, 2017.

[37] A. Chandrasekharan, M. Soeken, D. Große, and R. Drechsler, "Precise error determination of approximated components in sequential circuits with model checking," in *Proc. of DAC'16*. ACM, 2016, pp. 1–6.

[38] H. Jiang, C. Liu, L. Liu, F. Lombardi, and J. Han, "A review, classification, and comparative evaluation of approximate arithmetic circuits," *J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 4, 2017.

[39] H. Saadat, H. Bokhari, and S.Parameswaran, "Minimally biased multipliers for approximate integer and floating-point multiplication," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2623–2635, 2018.

[40] Z. Vasicek, "Formal methods for exact analysis of approximate circuits," *IEEE Access*, vol. 7, no. 1, pp. 177 309–177 331, 2019.

[41] S. Rehman, B. S. Prabakaran, W. El-Harouni, M. Shafique, and J. Henkel, *Heterogeneous Approximate Multipliers: Architectures and Design Methodologies*. Springer, 2019, pp. 45–66.

[42] I. Wali, M. Traiola, A. Virazel, P. Girard, M. Barbareschi, and A. Bosio, "Can we approximate the test of integrated circuits?" in *3rd Workshop On Approximate Computing (WAPCO)*, Jan. 2017, pp. 1–7.

[43] ——, "Towards approximation during test of integrated circuits," in *2017 IEEE 20th International Symposium on Design and Diagnostics of Electronic Circuits Systems (DDECS)*, April 2017, pp. 28–33.

[44] M. Traiola, A. Virazel, P. Girard, M. Barbareschi, and A. Bosio, "Towards digital circuit approximation by exploiting fault simulation," in *IEEE East-West Design Test Symposium*, Sep. 2017, pp. 1–7.

[45] ——, "Testing integrated circuits for approximate computing applications," in *4th Workshop On Approximate Computing*, 2018, pp. 1–7.

[46] ——, "Testing approximate digital circuits: Challenges and opportunities," in *2018 IEEE 19th Latin-American Test Symposium (LATS)*, March 2018, pp. 1–6.

[47] ——, "On the comparison of different atpg approaches for approximate integrated circuits," in *IEEE 21st International Symposium on Design and Diagnostics of Electronic Circuits Systems*, 2018, pp. 85–90.

[48] L. Anghel, M. Benabdenbi, A. Bosio, M. Traiola, and E. I. Vatajelu, "Test and reliability in approximate computing," *Journal of Electronic Testing*, vol. 34, no. 4, pp. 375–387, Aug 2018.

[49] M. Traiola, A. Virazel, P. Girard, M. Barbareschi, and A. Bosio, "Investigation of mean-error metrics for testing approximate integrated circuits," in *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems*, 2018, pp. 1–6.

[50] ——, "A test pattern generation technique for approximate circuits based on an ilp-formulated pattern selection procedure," *IEEE Transactions on Nanotechnology*, pp. 1–1, 2019.

[51] ——, "Maximizing yield for approximate integrated circuits," in *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2020.

[52] 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, Sept 2013.

[53] A. Chandrasekharan, S. Eggersglüß, D. Große, and R. Drechsler, "Approximation-aware testing for approximate circuits," in *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2018, pp. 239–244.

[54] A. Gebregiorgis and M. B. Tahoori, "Test pattern generation for approximate circuits based on boolean satisfiability," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2019, pp. 1028–1033.

[55] R. A. Frohwerk, "Signature analysis: a new digital field service method," 1977.

[56] V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan, "Analysis and characterization of inherent application resilience for approximate computing," in *2013 50th ACM/EDAC/IEEE Design Automation Conference (DAC)*, May 2013, pp. 1–9.

[57] A. Bosio, W. J. Perez-Holguin, and E. Sanchez, "Exploiting approximate computing to increase system lifetime," in *2019 IFIP/IEEE 27th Int. Conf. on Very Large Scale Integration (VLSI-SoC)*, 2019, pp. 311–316.

[58] V. Mrazek, R. Hrbacek, Z. Vasicek, and L. Sekanina, "Evoapprox8b: Library of approx 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.

[59] (2009) Nova. [Online]. Available: https://opencores.org/project/nova

[60] Xiph.org video test media. [Online]. Available: https://media.xiph.org/video/derf/

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.