Can we Approximate the Test of Integrated Circuits?
Imran Wali, Marcello Traiola, Arnaud Virazel, Patrick Girard, Mario Barbareschi, Alberto Bosio

To cite this version:

HAL Id: lirmm-02004418
https://hal-lirmm.ccsd.cnrs.fr/lirmm-02004418
Submitted on 1 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Can we Approximate the Test of Integrated Circuits?

Imran Wali1, Marcello Traiola1, Arnaud Virazel1, Patrick Girard1, Mario Barbareschi2, Alberto Bosio1
1LIRMM – University of Montpellier / CNRS – France
<name.lastname>@lirmm.fr
2Università Federico II – Naples
<name.lastname>@unina.it

Abstract—In the recent years Approximate Computing (AC) has emerged as new paradigm for energy efficient IC design. It addresses the problem of maintaining reliability and thus coping with run-time errors exploiting an acceptable amount of overheads in terms of area, performances and energy consumption. This work starts from the consideration that AC-based systems can intrinsically accept the presence of faulty hardware (i.e., hardware that can produce errors). The hardware-induced errors have to be analyzed to determine their propagation through the system layers and eventually determining their impact on the final application. In other words, an AC-based system does not need to be built using defect-free ICs. Under this assumption, we can relax test and reliability constraints of the manufactured ICs. One of the ways to achieve this goal is to test only for a subset of faults instead of targeting all possible faults. In this way, we can reduce the manufacturing cost since we eventually reduce the number test patterns and thus the test time. We call this approach Approximate Test. The main advantage is the fact that we do not need a prior knowledge of the workload (i.e., we are application independent). Therefore, the proposed approach can be applied to any kind of ICs, reducing the test time and increasing the yield. We present preliminary results on some simple case studies. The main goal is to show that by letting some faults undetected we can save test time without having a huge impact on the application quality.

Keywords—fault coverage; test pattern; approximate test; test generation; test complexity

I. INTRODUCTION

Today’s Integrated Circuits (ICs) are starting to reach the physical limits of CMOS technology. Among the multiple challenges arising from technology nodes lower the 20 nm, we can highlight the high leakage current (i.e., high static power consumption), reduced performance gain, reduced reliability, complex manufacturing process leading to low yield and complex testing process, and extremely costly masks [1]. In other words, ICs manufactured with the latest technology nodes are the less and less efficient (w.r.t. both performances and energy consumption) than forecasted by the Moore’s law. Moreover, the manufactured devices are becoming less and less reliable, meaning that errors can appear during the normal lifetime of a device with a higher probability than in previous technology nodes [2]. Fault tolerant mechanisms are therefore required to ensure the correct behavior of such device at the cost of extra area, power and timing overheads. Finally, process variations force the engineers to add extra guard bands (e.g., higher supply voltage or lower clock frequency than required under normal circumstances) to guarantee the correct functioning of manufactured devices.

In the recent years, the Approximate Computing (AC) paradigm has been emerged [2][3][4]. It addresses the problem of maintaining reliability and thus coping with run-time errors, exploiting an acceptable amount of overheads in terms of area, performances and energy consumption. AC is based on the intuitive observation that, while performing exact computation requires a high amount of resources, allowing selective approximation or occasional violation of the specification can provide gains in efficiency (i.e., less power consumption, less area, higher manufacturing yield) without significantly affecting the output quality [2][3][4].

This work starts from the consideration that AC-based systems can intrinsically accept the presence of faulty hardware (i.e., hardware that can produce errors) [5]. The hardware-induced errors have to be analyzed to determine their propagation through the system layers and eventually determining their impact on the final application. In other words, an AC-based system does not need to be built using defect-free ICs. In deed, AC-based systems can manage at higher-level the errors due to defectives ICs, or those errors simply do not significantly impact on the final applications. Under this assumption, we can relax test and reliability constraints of the manufactured ICs. One of the ways to achieve this goal is to test only for a subset of faults instead of targeting all possible faults. In this way, we can reduce the manufacturing cost since we eventually reduce the number of test pattern and thus the test time. In the literature, some interesting works have already been published so far, targeting the test generation for a subset of faults [6]–[7]. The main idea behind these works is the classification of faults as benign and malignant to be further exploited during the test generation. More in detail, the authors propose to generate all possible faults and then classifying them into two classes: 1) benign faults - those that cause no error or an acceptable amount of error, and 2) malignant faults - those faults that cause a significant deviation from acceptable behavior. The metric used for classifying faults into these two classes is the error magnitude [2], which is the difference between actual value (affected by the fault) and the golden value (fault-free).

The contribution of this work is to investigate an approach opposite w.r.t. the state-of-the-art. Instead to classify the faults according to the applied workload, we exploit a functional and a structural analysis to determine the most vulnerable circuit elements and thus generate the test patterns for these elements. We call this approach Approximate Test (AT). The main advantage is the fact that we do not need a prior knowledge of the workload (i.e., we are application independent). Therefore, the proposed approach can be applied to any kind of ICs reducing the test time and increasing the yield. In this paper,
we present preliminary results on some simple case studies. The results aim at looking at the impact of the AT on the test length and the fault coverage. Moreover, we also show the impact on the final application by using the well-known error probability (\(P_e\)) and the error magnitude (\(e\)) metrics [9]. The main goal is to show that by letting some faults undetected we can save test time without having a huge impact on the application quality.

The paper is organized as follows. Section II presents the flow of the proposed approach and gives details of each step. Experimental results are discussed in Section III. Finally, conclusions are given in Section IV.

II. THE PROPOSED APPROACH

The main idea of the proposed AT approach is to let some faults untested in order to speed up the test and to increase the overall yield. As already discussed, the risk of exploiting defective ICs to build AC-based systems is mitigated by the fact the AC can intrinsically accepts the presence of errors. However, the important point is that the impact of the untested faults on the final applications has to be in the acceptable regions (i.e., the final output quality is still acceptable by the user). The real challenge is therefore determining what are the faults that must be tested and what are those that can be ignored during the test application.

The straightforward approach for determining the targeted faults is act accordingly to the functionality of the circuit. The classical example is an arithmetic circuit where it is better to narrow down the \(e\) [2]. Unfortunately this approach cannot be adopted for all the kind of integrated circuits since it is not always easy to determine the most significant outputs. In deed the latter are clearly strictly related to the workload. To overcome the above issue, we propose to determine the targeted faults by using a structural analysis. In this way we want to be independent w.r.t. the circuit function. To validate our proposal, we developed a flow for applying both the approaches (i.e., functional and structural analysis).

The Fig. 1 describes the main steps of the proposed AT approach. The starting point is the circuit netlist (original design). The first step is the analysis of the netlist. Actually two types of analysis are performed: the Functional and the Structural. The goal of each analysis is to rank the circuit outputs based on their significance and susceptibility respectively. Then, in the second step, a fault list is generated depending on the output ranking previously computed. The fault list contains the fault that must be detected. The third step consists in running an ATPG with the circuit netlist and the generated fault list as inputs. The ATPG provides (in addition to the test patterns) the Fault Coverage (FC) and the Test Length (TL). The last step of the proposed flow computes the AC metrics (i.e. \(P_e\) and \(e\)). Those metrics are computed with the help of an exhaustive fault injection. Next sub-sections provide details on each step of the proposed flow.

A. Functional Analysis – Significance

The functional analysis aims at determining the targeted faults accordingly to the functionality of the circuit. As already discussed, this approach works very well for particular types of circuits. For example, when considering arithmetic ICs, outputs can be easily ranked depending on their weight. The Most Significant Bit (MSiB) of the output data word is the one having the most significance while the Low Significant Bit (LSiB) has the low significance on the computed result. Consequently, detecting faults affecting the fan-in cone of the MSiB guarantees the functionality with a minimum \(e\). Nevertheless, the significance analysis can only be computed on arithmetic ICs and not on random logic ICs. To be more general, the significance of the circuit outputs has to be
computed accordingly with the applied workload. The output of this step is the raking of the circuit outputs.

B. Structural Analysis – Susceptibility

Conversely to the functional analysis, we propose to determine the targeted faults by looking only at the circuit structure. In this case, the main idea is to analyze each primary output of the circuit to determine its susceptibility as described in [8]. The output susceptibility analysis is based on the fact that not all outputs of a circuit have the same susceptibility, which is a function of the number of nodes in its fan-in logic cone. It exploits the structural properties of the output fan-in cone to get their relative susceptibility estimates. In other words, we aim to identify the outputs more affected by the faults to determine its susceptibility.

Algorithm 1 shows the pseudo-code of the susceptibility analysis methodology. The algorithm starts by reading the place-and-route netlist of the design. Then, it forms groups \( F_j \) of all fan-in cells for each circuit output \( O_j \). Once groups are formed the weight \( W_j \) of each fan-in cone is calculated by adding together the weights of all cells in the corresponding fan-in cone group. According to the hypothesis that forms the basis of this methodology, cell weight is the number of inputs and outputs of that cell. Ranks are assigned to each output on the basis of their fan-in cone weight using a sort function shown in line 15 of Algorithm 1.

```
1 read(netlist);
2 // Group all fan-in cone cells together for each output node
3 foreach O_j do
4   F_j ← O_j.get_fanin();
5 end
6 // Get weight of fanin cone of each output
7 foreach O_j do
8   W_j ← W_j + O_j.get_fanin();
9 end
10 // Sort outputs on the basis of their fanin cone weight
11 sort(O_j,F_j,W_j);
```

Algorithm 1. Output susceptibility analysis

Because these figures we can infer that detecting the faults affecting the fan-in cone of \( O_1 \) significantly impacts the fault coverage while detecting the remaining faults slightly improves the fault coverage. As for the significance analysis, the output of this step is the raking of the circuit outputs from the Most Susceptible Bit (MSuB) to the Least Susceptible Bit (LSuB). As in the example of Fig. 2 they are \( O_1 \) and \( O_2 \) respectively.

C. Constrained Fault List Generation

This step takes in input the results of one of the two previous analyses (Functional or Structural). Please note that, if the circuit under consideration is a random logic circuit for which the workload is not known, only the structural analysis is possible. Having the circuit outputs ranked, we choose a subset of them that we want to cover during the AT. For each selected output, we add stuck-at-faults affecting its fan-in cone. The resulting fault list is thus composed of all the stuck-at faults observable at the selected outputs. Once again, the generated fault list only contains a sub-set of the total fault list.

Let us come back to the example shown in Fig. 2. For each output we determine the collapsed stuck-at-fault list.

```
1 read(netlist);
2 // Group all fan-in cone cells together for each output node
3 foreach O_j do
4   F_j ← O_j.get_fanin();
5 end
6 // Get weight of fanin cone of each output
7 foreach O_j do
8   W_j ← W_j + O_j.get_fanin();
9 end
10 // Sort outputs on the basis of their fanin cone weight
11 sort(O_j,F_j,W_j);
```

Algorithm 1. Output susceptibility analysis

The algorithm is further explained by its application to a simple example circuit shown in Fig. 2. The shaded regions mark the boundaries of the two output fan-in cones. The weight parameter \( (W_j) \) is given on the top of each gate. The fan-in cones weight \( (S_j) \) given on the right of corresponding output is found to be 14 and 10 for \( O_1 \) and \( O_2 \) respectively.

According to these figures we can infer that detecting the faults affecting the fan-in cone of \( O_1 \) significantly impacts the fault coverage while detecting the remaining faults slightly improves the fault coverage. As for the significance analysis, the output of this step is the raking of the circuit outputs from the Most Susceptible Bit (MSuB) to the Least Susceptible Bit (LSuB). As in the example of Fig. 2 they are \( O_1 \) and \( O_2 \) respectively.

Fig. 2. Application of the susceptibility analysis

According to these figures we can infer that detecting the faults affecting the fan-in cone of \( O_1 \) significantly impacts the fault coverage while detecting the remaining faults slightly improves the fault coverage. As for the significance analysis, the output of this step is the raking of the circuit outputs from the Most Susceptible Bit (MSuB) to the Least Susceptible Bit (LSuB). As in the example of Fig. 2 they are \( O_1 \) and \( O_2 \) respectively.

Fig. 2. Application of the susceptibility analysis

According to these figures we can infer that detecting the faults affecting the fan-in cone of \( O_1 \) significantly impacts the fault coverage while detecting the remaining faults slightly improves the fault coverage. As for the significance analysis, the output of this step is the raking of the circuit outputs from the Most Susceptible Bit (MSuB) to the Least Susceptible Bit (LSuB). As in the example of Fig. 2 they are \( O_1 \) and \( O_2 \) respectively.

According to these figures we can infer that detecting the faults affecting the fan-in cone of \( O_1 \) significantly impacts the fault coverage while detecting the remaining faults slightly improves the fault coverage. As for the significance analysis, the output of this step is the raking of the circuit outputs from the Most Susceptible Bit (MSuB) to the Least Susceptible Bit (LSuB). As in the example of Fig. 2 they are \( O_1 \) and \( O_2 \) respectively.

Fig. 3. Fault List generation from \( O_1 \)

Fig. 3. Fault List generation from \( O_1 \)

Fig. 3. Fault List generation from \( O_1 \)

Fig. 3. Fault List generation from \( O_1 \)

Fig. 3. Fault List generation from \( O_1 \)

Fig. 3. Fault List generation from \( O_1 \)

Fig. 3. Fault List generation from \( O_1 \)
also affecting P.

Fig. 4. Fault List generation from O2.

Fig. 4 graphically shows the collapsed fault list obtained by considering the fan-in cone of O2. It is composed of 7 stuck-at faults. We detailed this step since it is important to note the fan-out branch at the output of the invert C5. In the fault list from O1, the faults affecting the inverter have been considered. However, in O2 we must add the faults affecting the branch impacting on the O2 fan-in cone itself. Otherwise those faults will never be targeted during the Automatic Test Pattern Generation (ATPG).

D. ATPG

For this step of the AT flow, a commercial ATPG tool, with default options (i.e., fault dropping, dynamic compaction and random fill), is used to generate the test patterns. The Fault Coverage (FC) and the Test Length (TL) are two important outputs since they represent the effectiveness of the proposed AT compared to a deterministic one. We also save the patterns list to further compute the AC metrics.

Fig. 5. Test generation example.

Table I summarizes the test generation for the given example. Let us consider the first row. It reports the ATPG data executed by using the O1 fault list. The achieved fault coverage is 87.5% and 6 test patterns are generated. The column “detected faults” gives the detail per each fault list. It can be seen that targeting the O1 fault list will result in detecting for “free” 5 faults of the O2 fault list thanks to the random fill.

E. Exhaustive Fault Injection

Error Probability (Pε) and the Error Magnitude (ε) are metrics allowing the evaluation in the AC context. Pε represents the ratio of erroneous responses among the total responses that the circuit can produce. ε gives the hamming distance between the erroneous response and the golden one. Pε can be computed for any digital circuits while ε computation is only possible for arithmetic circuits or when the workload is known. Since we do not consider a particular workload, we compute both metrics by running an exhaustive fault injection campaign. All possible stuck-at-faults are injected (one after the other) and the exhaustive pattern list is simulated. It is worth to mention that this exhaustive fault injection can only be applied for small circuits (i.e., with a limited number of outputs). In this work is used only for validation, in the general case the workload has to be known in order to compute the metrics. Responses are then compared to the golden ones to compute Pε metric. ε metric is obtained in the same way but the hamming distance between the simulated fault injection and the golden response is obtained with the help of the output ranking computed during the functional analysis. Once again this is done only for arithmetic circuits since they do not require a workload to determine the weight of the outputs.

III. EXPERIMENTAL RESULTS

The AT flow has been validated on three case studies. All the circuits have been synthesized with a 45nm technology library 0. Table II reports the main characteristics of each circuit in terms of number of primary inputs (#. PI), primary outputs (#. POs), logic gates (#. Gates). All the case studies are purely combinational circuits. The first two are arithmetic circuits able to compute 4-bit sum, subtraction and logic operation. Moreover, the ALU2 can also compute comparisons between the two inputs (i.e., equals-to, greater-than and less-than). The last circuit is the c432 from the ISCAS’85 benchmarks suite [11]. We selected these benchmarks in order to have two cases for which both the functional and structural analyses can be applied, while for the last one, only the structural analysis is applied.
The maximum error of the ALU circuit is 0.199, indicating a probability of 0.199 to have a failure at the output (i.e., due to the untested faults) with a maximum deviation of 7. Therefore, by adding the remaining outputs we slightly increase the fault coverage, reaching 99.47% at the end. Considering the test length, we can notice that the number of test patterns varies from 19 up to 24. Thus, only considering one output, we can reduce the test length to 12.5%. Depending on the final system quality, the test engineer can therefore select the appropriate set of primary output leading to the desired trade-off between fault coverage, test length, $P$, and $\epsilon_{\text{max}}$.

Note that the last line in Table III represents the case of a standard deterministic test where all stuck-at-faults are considered during the test generation (i.e., the upper bound of the $FC$ and $TL$).

### B. ALU2

The Fig. 7 shows the block diagram of the ALU2 circuit, which can compute the standard arithmetic and boolean operation over the two 4-bit inputs. The circuit can compute the standard arithmetic and boolean operations. In the following two possible scenarios:

- **S1**: The logical outputs are more significant than the output data. Thus, the functional analysis will provide the rank $l$, $g$, $e$, $z_2$, $z_3$, $z_1$, and $z_0$.

- **S2**: The logical outputs are less significant than the output data. Thus, the functional analysis will provide the rank $z_0$, $z_2$, $z_1$, $z_0$, $l$, $g$, and $e$.

Let us verify S1 as the target scenario for our experiment, so that we can observe what happens when the output ranking is different for functional and structural analysis.

### Table IV. Functional Analysis: Scenario S1

<table>
<thead>
<tr>
<th>Outputs</th>
<th>FC(%)</th>
<th>TL</th>
<th>$P$</th>
<th>$\epsilon_{\text{avg}}$</th>
<th>$\epsilon_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76.19</td>
<td>27</td>
<td>0.287</td>
<td>1.551</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>76.62</td>
<td>29</td>
<td>0.257</td>
<td>1.328</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>77.16</td>
<td>31</td>
<td>0.250</td>
<td>1.271</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>94.83</td>
<td>55</td>
<td>0.192</td>
<td>0.536</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>97.52</td>
<td>62</td>
<td>0.145</td>
<td>0.245</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>98.28</td>
<td>66</td>
<td>0.082</td>
<td>0.082</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>100.00</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
On the other hand, if the complete fault list is generated the FC is 97.88%. So we simply loose 0.5% of fault coverage. Looking at the test complexity, we can reduce it by 4% if we consider only the MSuB output for the fault list generation.

### IV. Conclusion

The contribution of this work was to investigate an approach opposite w.r.t. the state-of-the-art. Instead to classify the faults according to the applied workload, we exploit a structural analysis to determine the most vulnerable circuit elements and thus generate test patterns for these elements. Preliminary results indicate that the resulting method, called *Approximate Test* can really lead to provide benefits in terms of test time reduction. Test engineer can select the desired trade-off between quality and test complexity. The proposed structural analysis methodology seems interesting and promising step toward making the AT applicable for any kind of circuits.

### References


---

**Table V. Structural Analysis: Scenario S2**

<table>
<thead>
<tr>
<th>Outputs</th>
<th>FC(%)</th>
<th>TL</th>
<th>P&lt;sub&gt;e&lt;/sub&gt;</th>
<th>ε&lt;sub&gt;avg&lt;/sub&gt;</th>
<th>ε&lt;sub&gt;max&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85.56</td>
<td>38</td>
<td>0.256</td>
<td>0.592</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>88.36</td>
<td>45</td>
<td>0.209</td>
<td>0.302</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>89.76</td>
<td>50</td>
<td>0.147</td>
<td>0.141</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>93.10</td>
<td>57</td>
<td>0.067</td>
<td>0.063</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>99.35</td>
<td>70</td>
<td>0.039</td>
<td>0.038</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>99.89</td>
<td>71</td>
<td>0.007</td>
<td>0.007</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>100.00</td>
<td>72</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Tables IV and 0 report the results obtained by using only the functional and the structural analysis respectively. We can immediately compare the tables. Let us consider the first row of both the cases. For the functional analysis, when considering only one output and its related fan-in cone set of stuck-at faults, we achieve 76% of FC. Conversely, when the output is provided by the structural analysis we can reach 85% of FC. There is a quite huge difference of 9% of FC. The P<sub>e</sub> and ε<sub>avg</sub>, ε<sub>max</sub> metrics also show a significant difference between the two cases. The ones associated to the structural analysis (see Table 0) are lower than those obtained by using the functional analysis. Especially the ε<sub>avg</sub> and ε<sub>max</sub> are reduced from 1.55 to 0.59 and from 15 down to 7 respectively. Which correspond to 2.6 times reduction in ε<sub>avg</sub> and a 2.1 times reduction in ε<sub>max</sub> when comparing our structural analysis with the functional one. Finally, both the analysis lead us a maximum reduction of about 45% of test length compared to the upper bound case (i.e. the last line of the tables where all stuck-at faults are considered during the test generation).

Similar to the first case study (see Section III.A), we can comment that depending on the final system quality, the test engineer can therefore select the appropriate analysis leading to the desired trade-off between fault coverage, test length, P<sub>e</sub> and ε metrics value.

However, a more interesting comment is about the differences between the structural and functional analysis. In this case, the structural analysis is even better than the functional one. Due to lack of results, we cannot claim that it is always true, but it is a non-intuitive result: only considering the structural information we can improve the coverage of the functionalities of the target circuit since both P<sub>e</sub> and ε are reduced.

**C. c432**

The last case study is the public available benchmark c432 [11]. As reported in Table II it has 36 primary inputs. Even if this number is not so high, it becomes to be not interesting to determine the P<sub>e</sub> by using the exhaustive fault injection (the complexity will rapidly increases). Indeed, this case study has been selected in order to put the study in a context close to real cases where the P<sub>e</sub> and ε can only be computed with a given workload. Thus, the Table VI does not provide any value for P<sub>e</sub> and ε.

Reported results are indeed interesting: by considering only one output (i.e., the MSuB), we can achieve the 97.32% of FC.

**Table VI. c432 Results**

<table>
<thead>
<tr>
<th>Outputs</th>
<th>FC(%)</th>
<th>TL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97.32%</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>97.77%</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>97.88%</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>97.88%</td>
<td>68</td>
</tr>
<tr>
<td>5</td>
<td>97.88%</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>97.88%</td>
<td>68</td>
</tr>
<tr>
<td>7</td>
<td>97.88%</td>
<td>68</td>
</tr>
</tbody>
</table>