Selective Test Response Collection for Low-Power Scan Testing with Well-Compressed Test Data

Dong Xiang  
School of Software,  
Tsinghua University,  
Beijing 100084, China  
dxiang@tsinghua.edu.cn

Zhen Chen  
Dept. of Computer Sci. and Techn.  
Tsinghua University  
Beijing 100084, China  
z-chen07@mails.tsinghua.edu.cn

Abstract

A new test application scheme is proposed for low-power scan testing, which is able to compress test data significantly. A combination of a scan architecture and an existent test compression scheme can compress test data even better. Test power can be reduced greatly based on the new test application scheme, according to which only a subset of scan flip-flops shifts a test vector or captures test responses in any clock cycle. Test response data can be another important problem. A new test response compaction scheme called selective test response collection is proposed to reduce test response volume. Selective test response collection combines with a structure-based test response compactor, according to which many test response data can be dropped. Experimental results show that the proposed test application scheme can efficiently reduce test power, compress test stimulus data, and compact test response data while test application cost can be well-controlled.

I. Introduction

Scan design makes test generation of a circuit to be that of a combinational one, which can obtain complete fault coverage and make test application time, test data volume, and test power prohibitively high. Test application time and test data volume can be reduced effectively by efficient test generation schemes, test compression techniques [14, 9, 17], and new scan architectures [2, 11, 16, 19, 20]. Scan flip-flop ordering, clock disabling, blocking logic insertion [18] can effectively reduce test power, especially clock disabling.

Test response compaction is another important issue to reduce test data volume. The test response compaction method in [15] is the first space only compaction scheme using error codes. The X-compact technique was proposed by Mitra, et al. in [13] based on the original Saluja-Karpovsky [15] scheme by increasing the input fanout. The method in [13] simultaneous errors and unknowns. Rajski, et al. [14] proposed a masking scheme to avoid unknowns by using a single bit extra data to control whether test responses of a scan chain are compacted into the test response compactor. The extra control data were generated by the decoder. Bhatacharyya, et al. [3] showed the possibility to design a test response compactor with a single output for a general design. However, the mapping logic looks non-trivial.

Many test compactors were proposed for BIST [4, 7, 5]. An optimization search based on genetic algorithms was proposed to compact test responses for BIST maximally in Ivanov, Tsuji, and Zorian [8]. Chakrabarty and Hayes [4] introduced a zero-aliasing test response compactor for BIST based on a graph coloring scheme. A hybrid compaction scheme was proposed in [5] for test response compaction by using a combination of blocking logic and an unknown blocking MISR.

Most of the previous test response compactor avoided aliasing by unmasking any of the fault effects propagated to scan flip-flops or POs. Fault effects of any detected fault can be propagated to multiple sites, however, it is enough if one of the sites can be observed. This idea is used to establish a new test response compactor. We would like to propose a new method that can significantly compress test stimulus data, and compact test responses, and effectively reduce shift and capture power. Test application time must also be comparable to that of multiple scan chain.

In the rest of the paper, the idea of selective test response collection is introduced in Section II. The method to establish the response compactor based on selective test response collection is presented in Section III. Experimental results are presented in Section IV, which is followed by the conclusions in Section V.

II. Selective Test Response Collection

Most of the previous methods for test data compression only considered stimulus data compression unlike EDT [14], which also presented techniques to compact test response data. The test response data volume can be even higher than that of stimulus data in many cases. Any good test compression scheme for stimulus data does not have any impact on test response data volume if the number of test vectors is the same. Test response data must also be well-compactored to reduce the total test data volume. In this paper, the test response compactor is combined with a new test compression scheme.
as presented in Fig. 1 and Fig. 2. The test compression scheme is a combination of scan forest [19] and any coding-based test compression scheme.

In our method, we specify all don’t cares of the test vectors when applying them to the circuits, therefore, no unknown signals are contained in the response data. Techniques like that in EDT [14] and the method of Mitra, et al. in [13] can be applied to eliminate the unknowns in the test response data by using a structural-analysis-based technique as presented in this paper. We would like to introduce a new scheme called selective test response collection to reduce test response data volume, which can be combined with other test-compression-based scan testing methods. The principle of selective test response collection is to reduce test response data volume by dropping the unnecessary fault effects, instead of collecting all responses like the conventional methods.

As stated in [19], multiple scan chains can be connected to the same XOR gate for test response compaction if any pair of scan flip-flops at the same level of the chains do not have any common combinational predecessor. This condition provides no aliasing fault for stuck-at fault testing. This condition is a little strong. The test response compactor based on selective test response collection can provide much weaker condition.

Two scan chains connected to the same XOR tree must be driven by the same clock signal. As shown in Fig. 3, all scan chains are partitioned into three subsets $G_1$, $G_2$, and $G_3$. Two faults $f_1$ and $f_2$ are detected by the test vector. Let fault effects of $f_1$ be propagated to $A$, $B$, $F$, and $G$, and fault effects of $f_2$ reach $C$, $E$, and $H$. It is not necessary to send test responses of scan chain subsets $G_2$ and $G_3$ back to the ATE. That is, only test responses captured by scan flip-flops in $G_1$ are enough. This can effectively reduce test response data volume.

Two scan chains connected to the same XOR tree must be driven by the same clock signal. As shown in Fig. 3, all scan chains are partitioned into three subsets $G_1$, $G_2$, and $G_3$. Two faults $f_1$ and $f_2$ are detected by the test vector. Let fault effects of $f_1$ be propagated to $A$, $B$, $F$, and $G$, and fault effects of $f_2$ reach $C$, $E$, and $H$. It is not necessary to send test responses of scan chain subsets $G_2$ and $G_3$ back to the ATE. That is, only test responses captured by scan flip-flops in $G_1$ are enough. This can effectively reduce test response data volume.

\[ TR = (\#out \cdot d + \#pos) \cdot r \cdot vec. \]  

The response data volume according to the proposed test application scheme can be estimated by Equation (1) as presented in Fig. 1. In Equation (1), $r$ is the number of scan chain subsets for the low-power test application scheme as shown in Fig. 1; $\#out$ is the output number of the test response compactor. All primary outputs receive test responses $r$ times for each test vector because the POs capture test responses for each subset of scan trees.

Our method captures test responses only once for the total test application period of a single test vector. The test response data volume can further be compacted to $TR_1$ as presented in Equation (2), where test responses are collected $r$ times at the scan-out signals and just once at the POs.

\[ TR_1 = (\#out \cdot d \cdot r + \#pos) \cdot vec. \]  

The test response data according to Equation (2) can still be large enough. It is not necessary to receive all test responses, some of which can be unnecessary for fault detection. Fault effects of a fault can be propagated to multiple locations, where only one of these fault effects is enough to detect the fault.

The selective test response collection scheme can be implemented by the revised architecture as presented in Fig. 1. This can be implemented easily by the current ATEs through selectively receiving the test response data from the scan-out signals. Usually, test response data can be further compacted.
The extra register may require skipping some states when one or more scan chain subsets do not need to capture test responses. One extra vector can be necessary to control test response compaction. How can we obtain the extra vector \((v_1, v_2, \ldots, v_n)\) for each test vector? How can we get the information for sensitized paths for each detected fault? We would like to answer the second question first. A new procedure is presented to construct the test response compactor, which is proposed in Section III.

For each fault covered by the test vector, the sensitization paths can be found easily, where the scan flip-flops to capture the test responses can be identified easily. It is not difficult to identify the subsets of scan chains, which has at least one scan flip-flop to capture fault effects for a fault. When fault effects of a single fault detected by the test vector are propagated to scan flip-flops in multiple subsets of scan chains, it is enough to observe its fault effect at a single site. Test responses captured by a subset of scan chains can be disregarded if the fault effects of all faults detected by the test vector are propagated to other subsets of scan chains. Therefore, test response data can be well-compactd.

The test application cost based on the proposed low-power test application scheme can be comparable to that of multiple scan chain architecture because our method uses the scan forest architecture. Each scan tree can contain a large number of scan chains.

III. Test Response Compactor Based on Selective Test Response Collection

The new test response compactor is constructed after the test set has been generated. It is not necessary to establish the scan forest before test generation. It is enough to have the scan flip-flop groups. The information for each test vector on the covered faults and the fault effect propagation paths for all the covered faults can also be found by the fault simulation process. The scan forest and the test response compactor are still constructed concurrently like [19]. Let the most scan tree depth and the group size be set to \(d\) and \(g\), respectively.

We would like to assign a coordination measure for any pair of scan flip-flops according to the generated test set. Selection of the smallest number of scan flip-flops to observe the fault effects of all detected faults can be formulated into a minimum cover problem. The minimum cover problem is a known NP-hard problem. We would like to find a heuristic procedure to get the minimum number of necessary observation scan flip-flops for each test vector: (1) Our method first finds the scan flip-flop \(c_1\), where the most number of faults detected by the test vector can be observed; (2) We select the next scan flip-flop \(c_2\), where the most number of faults, that are not observed at \(c_1\), can be observed from \(c_2\); (3) this process continues until all faults detected by the test vector can be observed from at least one flip-flop.

Our method would like to cluster scan flip-flops that are required observing test responses by the most test vectors into the same subset of scan chains, so that the test response data volume can be well-compactd and the test application cost can be reduced efficiently. Let us introduce a new measure to estimate the coordination measure \(p_v(i, j)\) between any pair of scan flip-flops \(i\) and \(j\) for test vector \(v\). For any test vector \(v\), \(p_v(i, j) = 1\) if scan flip-flops \(i\) and \(j\) must be observed for the test vector \(v\) as stated earlier in this subsection, otherwise, \(p_v(i, j) = 0\), which is obtained from the minimum covering solution.

Assume that each scan chain can be driven by the same clock signal. Let \((c_1, c_2, \ldots, c_g)\) be a scan flip-flop group, and \(V\) and \(S\) be the test set and the scan flip-flop set. Our method selects another group \(G_n\) of scan flip-flops \((c'_1, c'_2, \ldots, c'_g)\). Our method connects the scan flip-flops \(c'_1, c'_2, \ldots, c'_g\) to any of \(c_1, c_2, \ldots, c_g\). The summation of the coordination measures for each pair of scan flip-flops placed in the same scan chain as presented in Equation (3) should be maximized, when the connection overhead should be minimized. The proposed routing-aware scan testing scheme can effectively reduce the connection overhead when the coordination measure summation is not sacrificed. Let the scan flip-flop \(c'_j \in G_n\) be connected to the \(k\)th scan chain in the scan tree.

\[
M = \sum_{v \in V} \sum_{k=1}^{g} \sum_{c'_j \in G_n} \sum_{c_i \in C_k} p_v(i, j), \tag{3}
\]

where \(p_v(i, j)\) is the coordination measure of the test vector \(v\), and \(C_k\) is any one of the scan chain in the scan tree. Consider the third new scan flip-flop group \(G_n (c'_1, c'_2, \ldots, c'_g)\) can be connected to the same scan tree. Let \(C_1, C_2, \ldots, C_g\) be the scan chains in the first scan tree driven by the same scan-in pin. A scan flip-flop \(c'_j \in G_n\) is selected and connected to a scan chain if the coordination measure summation \(T\) as presented in Equation (3) can be maximized and the connection overhead is minimized. The above process continues until the first scan tree has been constructed. All other scan trees can be established similarly.

We would like to introduce a new technique to establish the test response compactor based on the selective test response collection scheme after the scan trees have been constructed. The principle to construct the test response compactor is to minimize the total number of scan chain subsets that are necessary to collect test responses for all test vectors. Assume each subset of scan chains are driven by the separate clock signals. We cluster scan chains, that capture test responses of the most common faults for the same test vectors, into the same subset. An XOR network is established for the selected subset of scan chains, which are driven by the same clock signal.

In order to construct the test response compactor, we would like to establish a chain connected graph \(CCG (N, E)\). A heuristic algorithm is proposed to construct
the test response compactor based on selective test response collection. The node set of the CCG includes all scan chains. Consider two scan chains $C_1 (c_1, c_2, \ldots, c_6)$ and $C_2 (c'_1, c'_2, \ldots, c'_6)$. Any pair of scan chains can be connected to the same XOR gate for the test response compactor if none of the scan flip-flop pairs $(c_1, c'_1), (c_2, c'_2), \ldots, (c_6, c'_6)$, one is selected as the necessary observation scan flip-flop and fault effect of the detected fault by any test vector is also observable at the other scan flip-flop. It is necessary to obtain the necessary observation flip-flop for each fault and keep the other scan flip-flops that receive the fault effect for the detected fault. The memory overhead for this is usually linear in most cases.

Each node $C$ in the CCG is assigned a weight, while each edge is also assigned another weight. The weight $w_1$ assigned to a node $C$ represents the corresponding summation of the coordination measures for any pair of scan flip-flops considering all test vectors. Estimation of $w_1$ for a scan chain is presented in Equation (4). The weight $w_2$ assigned to an edge $(C_1, C_2)$ in the CCG represents summation of the coordination measures between any pair of scan flip-flops in both scan chains $C_1$ and $C_2$.

\[ w_1 = \sum_{v \in V} \sum_{c_i, c_j \in C} p_v(c_i, c_j). \quad \text{(4)} \]

\[ w_2 = \sum_{v \in V} \sum_{c_i \in C_1, c_j \in C_2} p_v(c_i, c'_j). \quad \text{(5)} \]

Our method first selects the node $C_1$ in the CCG with the most weight. An edge with the most weight, which connects the two nodes $C_1$ and $C_2$, is chosen. The node with the most weight is selected when multiple edges connected to $C_1$ have the same most weight. The third node $C_3$, that is connected to both $C_1$ and $C_2$, is further selected similarly if the given number of scan chain in one subset of scan chain driven by the same clock signal still has not been chosen. The node $C_4$ must have the most weight summation for the edges $(C_1, C_3)$ and $(C_2, C_3)$. The node with the most weight is selected when multiple nodes connected to $C_1$ and $C_2$ have the most edge weight summation.

When the given number of scan chains still have not been selected, our method selects the fourth node $C_4$, which is connected to all $C_1, C_2$ and $C_3$ in the CCG. Our method chooses the node with the most weight summation for the edges $(C_4, C_1), (C_4, C_2), (C_4, C_3)$. Similarly, the node with the most weight is selected when multiple nodes have the same edge weight summation. The above process continues until the given number of scan chains have been selected or no scan chain can be connected to the XOR tree. This process continues until the set number of scan chains have been chosen for a single subset of scan chains driven by the same clock signal. The above process continues until all scan chains are selected.

A new test application scheme is proposed for the new test response compactor based on the selective test response collection, according to which only the necessary subsets of scan chains capture test responses. The scan chain subset, that does not need to capture test responses, does not need to refill the test vector. Therefore, test application time and test response data can be reduced significantly.

The scan chain subsets are swept for each test vector by recording the subsets of scan chains that must be activated to observe test responses. A subset of scan chains must capture test responses if at least one of the scan flip-flops in the scan chains must observe the fault effect of any fault detected by the test vector. The compacted test responses from the outputs of the test response compactor must be sent back to the ATE.

Each test vector is applied in the following way: The test vector is applied into a subset of scan chains when disabling all other scan chains. The process continues until all scan flip-flops have received the test vector. The proposed method checks each vector for the subsets of scan chains, which must capture and observe test responses for the vector. Each subset of scan chains, that must be observed for the test vector, captures test responses. The subset of scan chains is activated and all other scan chains are disabling when scan flip-flops in the scan chains subset capture test responses. Assume that at least one more subset of scan chains must capture test responses, the test data is again applied into this subset of scan chains when shifting out the test responses captured by the scan flip-flops in the scan chain subset. The process continues until all recorded subsets of scan chains have captured test responses, which are shifted out and sent to the ATE.

### IV. Experimental Results

The proposed scan testing scheme based on a combination of scan forest architecture and any coding based test compression scheme [17] has been implemented. The revised ATALANTA test generator by including a dynamic test compaction scheme is used to generate tests.

As shown in Table 1, columns with $TA$, $AP$, $CP$, and $PP$ represent test application time, average test power, capture power, and peak power reduction ratios compared to scan design with a multiple scan chain architecture. Columns with $r$, $group$, $size$ and $SI$ stand for the size of the extra register (the number of capture cycles per vector), the number of scan flip-flop groups, the size of the largest scan flip-flop groups, and the number of scan-in signals, respectively. Parameter $r$ is determined by the test time of the multiple scan chain architecture with the same number of scan-in signals by considering a trade-off between the test power and test application time. Column vec represents the numbers of test vectors. Column area presents the area overheads.
Table 2: Performance on stimulus test data compression

<table>
<thead>
<tr>
<th>circuit</th>
<th>Full Scan</th>
<th>[17]</th>
<th>no SE</th>
<th>[19]</th>
<th>proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>s38532</td>
<td>15603</td>
<td>15603</td>
<td>183</td>
<td>393</td>
<td>435</td>
</tr>
<tr>
<td>b19</td>
<td>249600</td>
<td>249600</td>
<td>217</td>
<td>217</td>
<td>217</td>
</tr>
<tr>
<td>des</td>
<td>18360</td>
<td>18360</td>
<td>183</td>
<td>183</td>
<td>183</td>
</tr>
<tr>
<td>vga</td>
<td>931064</td>
<td>931064</td>
<td>931064</td>
<td>931064</td>
<td>931064</td>
</tr>
</tbody>
</table>

Table 3: Performance of response compaction

<table>
<thead>
<tr>
<th>circuit</th>
<th>Full scan</th>
<th>no SE</th>
<th>proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>s38532</td>
<td>15603</td>
<td>15603</td>
<td>435</td>
</tr>
<tr>
<td>b19</td>
<td>249600</td>
<td>249600</td>
<td>217</td>
</tr>
<tr>
<td>des</td>
<td>18360</td>
<td>18360</td>
<td>183</td>
</tr>
<tr>
<td>vga</td>
<td>931064</td>
<td>931064</td>
<td>931064</td>
</tr>
</tbody>
</table>

Table 4 presents comparison on test data volume of the proposed method with the previous methods by Li and Chakrabarty [11], the selective encoding scheme [17], the EDT scheme [14], the test compression scheme in Krishna, et al. [9], and the method in [12]. We present results of a combination of scan forest and selective encoding (proposed). The last column in Table 4 shows results of the proposed method with a commercial ATPG tool.

V. Conclusions

A combination of the scan architecture and a test data compression scheme can more effectively compress test data in low-power scan testing environment. Test response data were compacted effectively by using the structure information of the circuit. A new scheme...
called selective test response collection was proposed to compact test response data further based on a simple graph theoretic model. Experimental results showed that the proposed method can reduce peak test power and capture test power to less than 10% compared to the original ones and compress test stimulus data close to 600X for the available benchmark circuits. Test response data can be compacted up to 233X, while the selective test response collection scheme can further compact close to 6X with zero-aliasing compared to the structural-analysis-based test response compactor [19].

**References**


