A Unified Solution to Scan Test Volume, Time, and Power Minimization

Zhen Chen  
*Tsinghua University, Beijing, z-chen07@mails.tsinghua.edu.cn*

Sharad C. Seth  
*University of Nebraska - Lincoln, seth@cse.unl.edu*

Dong Xiang  
*Tsinghua University, dxiang@tsinghua.edu.cn*

Bhargab B. Bhattacharya  
*Indian Statistical Institute, Kolkata, India, bhargab@isical.ac.in*
A Unified Solution to Scan Test Volume, Time, and Power Minimization

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

Sharad Seth  
Computer Science and Engineering,  
University of Nebraska-Lincoln,  
Lincoln NE 68588-0115, U.S.A.  
seth@cse.unl.edu

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

Bhargab B. Bhattacharya  
Indian Statistical Institute,  
Kolkata, India  
bhargab@isical.ac.in

Abstract — The double-tree scan-path architecture, originally proposed for low test power, is adapted to simultaneously reduce the test application time and test data volume under external testing. Experimental results show significant performance improvements over other existing scan architectures.

Keywords: Test power minimization, Test time reduction, Test data reduction, Nonlinear scan

I. INTRODUCTION

Recent trends in manufacturing have spurred research on reduction of test power, test volume, and test-application time. Excessive test power can lead to unreliable operation, yield loss, device burnout, or reduced battery life, whereas increased test volume and test time contribute to the cost of testing.

For STUMPS-like scan architecture [1], with multiple parallel scan chains, it can be shown that as the design size is increased, the test power per cycle grows linearly, while test energy, test volume, and test time all grow at an even higher rate [2]. As the traditional scan schemes do not scale up well for today’s complex chips, containing these showstopper trends through other means has been an active area of research in recent years.

A large body of literature exists on reducing test data volume. Since lower test data volume leads to shorter test time, reducing test data volume is critical to low cost testing. In order to reduce the numbers of test vectors, static and dynamic methods are proposed to exploit the large fraction of don’t care bits [3, 6] in the ATPG test cubes. In order to reduce the data volume stored in the tester memory, various compression techniques have been proposed. These fall broadly into three categories [7]; (1) Code-based schemes [8, 7], (2) Linear-decompression-based schemes [2, 9, 10, 11], and (3) Broadcast-scan-based schemes [12]. Commercial tools based on the second and third kinds of methods are available since they are more efficient in compressing test vectors for industrial circuits with a large fraction of don’t care bits. Broadcast-scan-based methods have the advantage of incorporating the constraints for the decompressor into ATPG to produce encodable test cubes, over linear-decompression-based schemes. However, broadcast scan has less encoding flexibility because certain groups of scan cells receive identical values. In this paper, we use a broadcast-scan-based scheme that relaxes this constraint.

Further, test power is also an important metric of test performance [13]. Low power methods are mainly DFT-based [14] or pattern-based [15]. However, many data compression schemes are not compatible with the existing low power techniques. For example, For example, Illinois Scan [11], reconﬁgurable Illinois scan [4], dynamically reconﬁgurable shared scan architecture [5], and EDT [2] can efficiently reduce test data volume and test time, but have little effect on test power. Indeed, EDT effectively randomizes the don’t care bits in the test cubes, thereby resulting in scan-shift switching activity that is close to 50%. A low-power extension of EDT [16] intelligently maps scan slices to self-loop states of the ring generator to minimize transitions during scan shifts. However, the resulting improvement in test power is achieved at some sacrifice in data-compression efﬁciency.

In this paper, we present a unified solution to improve the three metrics simultaneously, based on the nonlinear, double-tree scan (DTS) architecture [17]. That work focused on scan-load power consumption, reducing both the peak load power and energy by the factor $N/\log N$, as compared to a linear scan chain of length $N$. DTS does not address the capture power issue but is compatible with schemes, e.g. [18], proposed to solve it. The DTS architecture, however, also offers opportunities for dynamically changing the mapping from input test-bit stream to the scan cells and loading the same input stream in multiple ways. These features allow us to weaken the constraint of broadcast-based compression and achieve significant reductions in test volume and test time, on top of the inherent power savings of DTS.

The rest of the paper is organized as follows. Section II presents the background and basic theory for our proposed method. Section III gives the overall idea. The test scheme and detailed strategies are proposed in Section IV. Experimental result are given in Section V. Section VI concludes the paper.

II. BACKGROUND

In this section we briefly describe the prior work that forms the basis for this paper. Specifically, we cover the salient aspects of the following: the double-tree scan architecture [17], and the no-common-successor relation between
scan cells that is used in the clustering algorithm for assigning cells to different scan chains.

A. Double-Tree Scan

A complete binary tree of level \( k \) (considering the root at level 0) consists of \( 2^k \) leaf nodes and \( (2^k - 1) \) internal nodes. The proposed scan structure resembles two complete \( k \)-level binary trees whose leaf nodes are merged pair-wise. Thus, a full double-tree DTS\((k)\) consists of \( N = (2^k - 1 + 2^k + 2^k - 1) = 3 \times 2^k - 2 \) nodes. Each node of the tree represents a scan flip-flop. All edges in the tree are directed from top to bottom. A directed edge \((i, j)\) in the DTS indicates that \( Q(i) \), i.e., the Q-output of the flip-flop \( i \), drives \( D(j) \), the D-input of the flip-flop \( j \). For each node with in-degree 2 (i.e., a merge node) in the bottom half of the DTS, a 2\(\rightarrow\)1 MUX is needed to select the predecessor flip-flop during the scan-load operation. Figure 1(a) shows a full DTS\((3)\) with 22 nodes. By pruning a full DTS, it is possible to obtain a partial DTS with any number of nodes; Figure 1(b) shows an 18-node partial DTS obtained by pruning DTS\((3)\). The forking (merging) nodes require the addition of a 1\(\rightarrow\)2 DEMUX (2\(\rightarrow\)1 MUX) to the corresponding scan cells (see Fig. 2), while the middle part can be built from standard scan cells. Both trees are balanced because each path from the source node to the sink node has the same length.

DTS can serve as a direct replacement of a scan chain by connecting the topmost (source) node to the scan-in signal and the bottommost (sink) node to the scan-out signal. The full DTS\((k)\) has \( 2^k \) overlapping scan paths, each of length \( (2k + 1) \), from the source to the sink. The control unit to load and unload the DTS repeatedly selects a source-to-sink scan-shift path in each clock cycle, so that, externally the DTS is indistinguishable from a serial scan chain of length \( N \), where \( N \) is the number of FFs in the DTS. However, in each clock cycle, only \( O(\log N) \) FFs are enabled, thus providing \( O(N/\log N) \) improvement in test power per cycle. Note that unlike the schemes that use linear chains and must minimize the chain length to minimize power consumption during scan loads, we can afford to implement larger DTS structures because of its logarithmic scan-shift depth. In other words, the design tradeoff between the number of chains and chain-length is very different for, and favorable to, DTS-implemented chains than linear chains.

A particularly simple control scheme for DTS is the breadth-first load [17], which involves selection of source-to-sink paths in a cyclic fashion until the tree is fully loaded. This idea can be realized by a counter in a centralized control, where each counter bit connects to the enable signals (\( E(i) \) in Figure 2) of scan cells at the same level in the top part of the DTS. The idea can also be realized by a distributed control by adding a toggling branch FF (see Fig. 3) to each scan cell in the top part to route the enable signal along the selected path. The branching FF BR[i] determines whether the next scan-shift path going through this cell will select the left or the right node of the DTS. The branch FFs can as well be inserted in the bottom part and the enable signal is routed from bottom to top so as to avoid the race condition between the data and the enable signal that might otherwise occur in the alternate design [19].

Results of our preliminary investigations on achievable
power savings with DTS and its impact on the area and routing of a design became recently available [19]. A fully synchronous test chip consisting of a sinc³ decimation filter with 188 flip-flops and 8000 gates was designed and simulated while clocking in a random test pattern in both the conventional scan path and the DTS scan path. The total number of logic level toggles in the DTS design decreased by a factor of 8 over the standard scan.

For measuring the hardware overhead of the DTS architecture, a RISC CPU, compatible with TI MSP 430 microcontroller core, and several ITC99 benchmarks were synthesized from the RTL-VHDL, in four different versions: without scan, with standard scan, with DTS using centralized control, and with DTS using distributed control. The designs were mapped to the IBM 0.18 micron process. The results show a centralized DTS area overhead of 16% over standard scan for most circuits. The distributed control overhead was another 15% on top of that. In all cases, it was possible to use the same standard cell utilization (over 90%) for all versions of the design, i.e. routing was not an issue.

The area overhead for both the centralized and distributed control schemes can be reduced significantly by including more than one scan FF per DTS node, with commensurate reduction in power savings. The area overhead of the distributed scheme can be further reduced by sharing of the branch FFs between two or more scan FFs at the same level of the DTS tree.

B. No-Common-Successor Relation

The effectiveness of the broadcast test mode in ILS depends strongly on the quality of the generated test patterns. Another approach [20], which we follow here, attempts to break the correlation between the scan cells based on the circuit topology. It makes use of the no-common-successor (NCS) binary relation between scan cells. Two scan cells are related by NCS if they cannot reach a common node in the circuit through combinational paths. Other researchers have used the complementary relation in defining a reconﬁgurable scan-chain architecture [21]. The usefulness of the NCS relation in minimizing correlation arises from the following lemma [20].

**Lemma 1** In a full scan circuit, if two scan FFs have the NCS relation, at most one of them needs to be assigned a binary value, in the generated test cube, to detect a single stuck at fault.

Because of the lemma, it is clear that two scan FFs in NCS relation can be assigned the same binary value without any loss of fault coverage.

III. Main Idea

Unlike the linear scan chain, there are multiple paths from the source to the sink in the DTS. We exploit this property, in the context of loading of parallel scan chains, where each chain is replaced by a DTS structure. Specifically, during the broadcast mode of Illinois Scan, this property allows the broadcast pattern to be permuted (see Section IV.A for more details) as it is loaded into different scan chains, thus largely overcoming the two major shortcomings of Illinois Scan: (1) identical bits loaded into the same bit positions of different scan chains, and (2) strong dependence of performance on scan-chain configuration. Further, simply by changing the relative permutations of load sequence between two chains, the same pattern can be repeatedly applied to provide additional fault coverage with negligible increase in the test-stimuli volume: with \( k \)-bits/scan-chain overhead, it is possible to specify any one of \( 2^k \) different permutations. Our results show that even two permutations of a pattern provide a dramatic increase in the achievable fault coverage.

IV. Our Scheme

Our scheme depends on modifying both the DTS and ILS in significant ways that are described in this section.

A. Input Permutations with DTS

Earlier, we indicated that the DTS with \( N \) nodes can be loaded in a breadth-first fashion by cycling through its paths for \( N \) cycles [17]. The specific order in which the paths are cycled determines the mapping from index of the bit in the input bit stream to the index of the FF in which the bit is loaded. Therefore, simply by changing the order in which the paths are cycled, we can load a pattern with different permuted bits.

Consider the breadth-first loading of partial DTS of six nodes shown in Figure 4 for illustration. This DTS has only two paths, that can be selected by specifying one path bit. Let us assume that the path bit value 0 denotes the left shift path and 1 denotes the right shift path. If we assume the initial path bit value to be 0 and the input bit-stream to be \(<b_5, b_4, b_3, b_2, b_1, b_0>\), where \( b_0 \) arrives first, then the resulting mapping will be as shown in Table 1 under the leftmost-first. Similarly, when the path bit value is initialized to 1, the resulting mapping will be as shown under the rightmost-first in Table 1. It is noteworthy that except for the source and sink FFs, all other FFs receive different bits in the input stream. For the leftmost-first and rightmost-first loads, this observation is generally true for any balanced DTS structure. We call a mapping from input bit-stream to scan FFs an input permutation. For the example DTS of six nodes, we can realize two different input permutations. Similarly, in a full DTS(\( k \)), there are \( 2^k \) different paths and we can realize \( 2^k \) input permutations, simply by initializing the path counter to different states, and another \( 2^k \) different permutations by running the counter in reverse.

This feature of DTS is potentially very useful in test-volume reduction because it allows reuse of a given pattern in \( 2^k \) different ways by specifying only \( k \) bits, where \( k \) is logarithmic in the size of the DTS. For example, for DTS(10) with 3,070 FFs, 1024 different permutations can be realized by specifying just 10 bits.

<table>
<thead>
<tr>
<th>Table 1: Leftmost-first and rightmost-first loading of the DTS shown in Figure 4.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan FF</td>
</tr>
<tr>
<td>Leftmost-first</td>
</tr>
<tr>
<td>Rightmost-first</td>
</tr>
</tbody>
</table>

B. Scan-Cell Clustering

Because of the lemma in Section II, it is clear that two scan FFs in NCS relation can be assigned the same binary
value without any loss of fault coverage. Although the NCS relation is exact and very useful, using the strict NCS condition for each group is prohibitive for some circuit with large input cones and output cones. Instead, we relax the NCS condition to construct scan chains. For the source nodes and sink nodes, we use the NCS condition to break the correlation between them. For nodes at other levels, the NCS condition needs not hold strictly because the correlation between them can be broken by loading DTS with different permutations.

The scan-cell clustering algorithm works on the NCS relation between scan FFs, represented as a graph. The goal of the algorithm is to form groups of $m$ scan FFs, where $m$ is the number of DTS chains, such that the $m$ flip-flops in a group are distributed among the $m$ DTS chains by placing one FF at the same level of each DTS tree. To form each group, the following steps are taken:

1. Initialize a new group $G$ to a null set.
2. Find the as-yet unprocessed node in the NCS graph with the highest degree, and use it as the seed of a new cluster $S$.
3. Repeatedly choose a node satisfying the NCS condition with all elements in $S$, until no such node can be found or the size of $S$ is $m$. The resulting set $S$ defines a cluster.
4. Put $S$ into $G$. If the size of $G$ is smaller than $m$, form additional clusters and put them into $G$ until $G$ becomes a group of size $m$.

After the groups are formed, we first choose two groups with the fewest clusters in them, and assign one group to all the source nodes the other group to all the sink nodes of the $m$ DTS chains. Next, the DTS nodes connected to the source and the sink nodes are assigned in the same fashion. Proceeding in this way, all the interior levels of the DTS are filled from the boundary to the middle of the DTS chains. For the scan flip-flops in a group, it does not matter which specific chains they are assigned to, keeping all the degrees of freedom for placement and routing.

C. Test Generation and Application

We assume that we have $m$ balanced scan chains, each of length $l$ and each implemented as a full or balanced-partial DTS (see Figure 1(a) and 1(b)). Further, we assume that no output data compression is being performed so that there are $m$ input channels and $m$ output channels, for a total of $2m$ scan channels. As the cost of ATE is directly proportional to the number of scan channels, we assume $m$ to be small, in the range of 2 to 32.

The test generation procedure is as follows:

1. Partition the scan cells into groups, using the clustering algorithm described in the Section IV.B. Then, modify the test bench to connect the scan cells in the same group to a common input.
2. Run the constrained ATPG on the circuit with the constraint that the scan flip-flops in the same group must be assigned the same value in the test generation process. For each pattern generated, use Step 3 to do fault simulation.
3. Fault-simulate with different permutations. (In the example below, we assume there are 8 chains.)
   a. The test pattern with the loading permutation 00000000 (i.e. leftmost-first load).
   b. The test pattern in which the loading of the first half of the chains is changed, that is, the loading permutation is 11110000.
   c. The test pattern in which the loading of the last half of the chains is changed, that is, the loading permutation is 00011111.

If additional faults are detected in any of the above sub-steps, the permutation is recorded.

4. Run another constrained ATPG, in which grouping of scan FFs is changed from Step (2) to realize a different permutation of the original pattern, which can be described with reference to Figure 5. We note that, except for the groups of source nodes and sink nodes, the remaining groups of $m$ nodes in a DTS can be paired according the permuted bit values they receive in the leftmost-first and rightmost-first loads, e.g. the group of nodes with hexagon shape (flip-flops 2 and 6) and the group of nodes with square shape (flip-flops 3 and 7) form a pair. If the left tree loads data in leftmost-first while the right tree in rightmost-first, the group pairs in Fig. 5 become (2, 7) and (3, 6) from the original pairs (2, 6) and (3, 7). After running the ATPG, we repeat the three sub-steps of Step 3, for the pattern obtained.

Let us give more explanation of step 4. In Figure 5, let us assume that scan cells marked with identical shape are in the same group, i.e. cells 1 and 5, 2 and 6, 3 and 7, 4 and 8. Assume, the detection of a fault requires the values of scan cells 2, 3, 6, 7 to be 1, 0, 0, X. In the constrained ATPG in Step 2, the fault is redundant because scan cell 2 and scan cell 6 must be the same value. But this fault can be detected in the Step 4. In Step 4, scan cell 2 and 6 can have the different values if one tree loads leftmost-first and the other tree loads rightmost-first. In essence, Step 4 is also the realization of using a different permutations to improve the fault coverage.

Test application under the described scheme requires that when a test pattern is repeated, the repeat code also includes an $m$-bit loading permutation vector that is used for initialization of each scan chain before the test pattern is loaded.

V. Experimental Results

The proposed scheme was implemented and run on ISCAS89 scan circuits to assess its performance. We use Atlanta as the basic ATPG tool to realize constrained ATPG.
power reductions are different according to the number of
scan chains, but, in all cases, the test cost reduction is sig-
ificant compared to full scan. So the number of scan chains
can be used as a tradeoff among the three test metrics.

We also compare our test-data volume results against
EDT (Table X of reference [2]) for the first six of the seven
ISCAS’89 circuits shown in Table 2 above. The ratios of
EDT test volume to ours for the six circuits are 1.80, 1.56,
1.19, 0.97, 1.19, and 0.98, respectively. That is, in four of
the six cases we achieve substantially better results and in
two cases EDT is slightly better.

VI. Conclusions

In this paper we show that in combination with an Illinois-
Scan like test application scheme, DTS can yield significant
improvements in three important metrics of interest to test
engineers, related to the data volume, time and power con-
sumption of tests. To achieve these benefits, we outlined
control schemes for DTS, described a clustering algorithm
to distribute scan cells between parallel scan chains, and
proposed a test-application scheme that exploits both the
ability of DTS to apply permuted test patterns and the re-
peat function of the ATE. Compared to linear scan, the DTS
scan cells include incrementally more logic and extra fanin
or fanout. We believe, the extra design cost of scan cells are
well worth the price for the benefits that follow from its use.

References

for VLSI Pseudorandom Techniques,” John Wiley and Sons,
bedded deterministic test,” IEEE Transactions on Computer-
gorithm for combinational circuits,” in IEEE/ACM Inter-
national Conference on Computer-Aided Design (ICCAD),
for reducing test time and test data volume in Illinois Scan
Architecture based designs,” in VLSI Test Symposium, 2002,
pp. 9-15.
R. Kapur, and T. W. Williams, “Changing the scan enable
during shift,” in VLSI Test Symposium, 2004, pp. 73-78.
paction for automatic test pattern generation,” IEEE Trans-
actions on Computer-Aided Design of Integrated Circuits and
[8] A. Jas, J. Ghosh-Dastidar, and N. A. Touba, “Scan vec-
tor compression/decompression using statistical coding,” in
IEEE VLSI Test Symposium (VTS), 1999, pp. 114-120.
determination for test volume and time reduction through
unified test pattern compaction and compression,” in IEEE
VLSI Test Symposium, 2003, pp. 113-118.
Farnsworth, and D. Wheater, “A SmartBIST variant with
guaranteed encoding,” in Proceedings 10th Asian Test Sym-
<table>
<thead>
<tr>
<th>name</th>
<th>Serial full scan</th>
<th>Performance-gain (ratio) over serial full scan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vec</td>
<td>data</td>
</tr>
<tr>
<td>s5378</td>
<td>251</td>
<td>44929</td>
</tr>
<tr>
<td>s9234.1</td>
<td>568</td>
<td>77648</td>
</tr>
<tr>
<td>s3207.1</td>
<td>436</td>
<td>278168</td>
</tr>
<tr>
<td>s5850.1</td>
<td>472</td>
<td>252048</td>
</tr>
<tr>
<td>s38417</td>
<td>886</td>
<td>1449496</td>
</tr>
<tr>
<td>s58544.1</td>
<td>658</td>
<td>938308</td>
</tr>
<tr>
<td>s35932</td>
<td>68</td>
<td>117504</td>
</tr>
</tbody>
</table>

Table 2: Comparison of ILS and our method on test cost.

<table>
<thead>
<tr>
<th>name</th>
<th>coverage of full scan</th>
<th>#sc</th>
<th>coverage of broadcast</th>
<th>coverage of permutations</th>
<th>Last stage</th>
<th>data</th>
<th>cycle</th>
<th>power</th>
<th>rpt</th>
<th>cbits</th>
<th>Performance-gain (ratio) over full scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>s5378</td>
<td>99.131</td>
<td>2</td>
<td>252</td>
<td>99.16</td>
<td>218</td>
<td>99.16</td>
<td>19620</td>
<td>23202</td>
<td>10</td>
<td>1.17</td>
<td>0.003</td>
</tr>
<tr>
<td>s9234.1</td>
<td>93.475</td>
<td>2</td>
<td>236</td>
<td>97.784</td>
<td>199</td>
<td>99.14</td>
<td>8955</td>
<td>11081</td>
<td>17</td>
<td>1.21</td>
<td>0.017</td>
</tr>
<tr>
<td>s3207.1</td>
<td>93.767</td>
<td>4</td>
<td>317</td>
<td>90.737</td>
<td>282</td>
<td>93.67</td>
<td>14946</td>
<td>17707</td>
<td>22</td>
<td>1.31</td>
<td>0.009</td>
</tr>
<tr>
<td>s13207.1</td>
<td>98.462</td>
<td>8</td>
<td>260</td>
<td>83.403</td>
<td>236</td>
<td>91.41</td>
<td>6102</td>
<td>7523</td>
<td>33</td>
<td>1.19</td>
<td>0.043</td>
</tr>
<tr>
<td>s38417</td>
<td>99.475</td>
<td>4</td>
<td>408</td>
<td>98.451</td>
<td>374</td>
<td>98.43</td>
<td>59840</td>
<td>76148</td>
<td>25</td>
<td>1.26</td>
<td>0.006</td>
</tr>
<tr>
<td>s58544.1</td>
<td>96.682</td>
<td>8</td>
<td>495</td>
<td>95.631</td>
<td>393</td>
<td>96.68</td>
<td>25662</td>
<td>60475</td>
<td>31</td>
<td>1.14</td>
<td>0.004</td>
</tr>
<tr>
<td>s35932</td>
<td>99.131</td>
<td>8</td>
<td>404</td>
<td>95.855</td>
<td>462</td>
<td>96.67</td>
<td>2416</td>
<td>28488</td>
<td>60</td>
<td>1.14</td>
<td>0.014</td>
</tr>
<tr>
<td>s9234.1</td>
<td>99.475</td>
<td>16</td>
<td>282</td>
<td>96.934</td>
<td>223</td>
<td>97.22</td>
<td>8920</td>
<td>12247</td>
<td>60</td>
<td>1.34</td>
<td>0.016</td>
</tr>
<tr>
<td>s38417</td>
<td>96.682</td>
<td>16</td>
<td>246</td>
<td>93.14</td>
<td>296</td>
<td>96.67</td>
<td>10064</td>
<td>12688</td>
<td>92</td>
<td>1.22</td>
<td>0.079</td>
</tr>
<tr>
<td>s58544.1</td>
<td>99.475</td>
<td>32</td>
<td>347</td>
<td>84.228</td>
<td>306</td>
<td>99.46</td>
<td>26522</td>
<td>34277</td>
<td>129</td>
<td>1.28</td>
<td>0.086</td>
</tr>
<tr>
<td>s35932</td>
<td>98.9</td>
<td>32</td>
<td>389</td>
<td>96.837</td>
<td>24</td>
<td>89.84</td>
<td>1296</td>
<td>3597</td>
<td>174</td>
<td>1.38</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Table 3: Performance under different number of scan chains.


