A Novel Scan Architecture for Power-Efficient, Rapid Test[*]

Ozgur Sinanoglu and Alex Orailoglu

Computer Science and Engineering Department
University of California, San Diego
La Jolla, CA 92093

{ozgur, alex}@cs.ucsd.edu

Abstract

Scan-based testing methodologies remedy the testability problem of sequential circuits; yet they suffer from prolonged test time and excessive test power due to numerous shift operations. The high density of the unspecified bits in test data enables the utilization of the test response data captured in the scan chain for the generation of the subsequent test stimulus, thus reducing both test time and test data volume. The proposed scan-based test scheme accesses only a subset of scan cells for loading the subsequent test stimulus while freezing the remaining scan cells with the response data captured, thus decreasing the scan chain transitions during shift operations. The experimental results confirm the significant reductions in test application time, test data volume and test power achieved by the proposed scan-based testing methodology.

1 Introduction

Even though testability of sequential circuits is significantly improved through the use of scan-based test methodologies, the shift operations employed for loading and observing the test data deteriorate test application time, consisting approximately of the product of the number of scan cells and test vectors in cycles in a scan-based environment. As each manufactured chip needs to be tested, prolonged test time incurs increased overall design cost.

In addition to test time prolongation, scan-based testing suffers from increased test power; during shift operations, the frequent transitions in the scan chain reflect into rippling at the circuit lines unnecessarily, hence resulting in increased power dissipation. The consequent overheating of the chip during test may in turn damage the chip. Test power should evidently be reduced to ensure reliable chip tests.

The justification of only a small subset of the primary inputs of a circuit for fault excitation and observation typically results in the generation of test sets consisting mostly of unspecified bits. These *don't care* bits should be exploited so as to reduce test time; appropriate fill-in of these bits may result in creating correlation in test data which may in turn be utilized to eliminate several shift operations that would otherwise be needed to load and observe test data.

Test power in a scan-based environment can be kept below certain thresholds through limiting the rippling that occurs in the scan chain. In this paper we propose a novel scan-based test scheme wherein only a subset of scan cells are loaded and observed whereas the remaining cells are frozen with the test response data captured. Scan chain rippling is restricted to occur only inside the scan chain fragment that is controlled and observed, reducing test power. The *don't care* bits in the test data are exploited as the corresponding scan cells need not be controlled; instead, the response data in these cells constitute the subsequent test stimulus data, implicitly setting the *don't care* bits. Control and observation of a small subset of scan cells not only limits the test power dissipated but furthermore reduces test time as well since the effective length of the scan chain is decreased. The scan cells to be accessed via the scan-in and scan-out pins are chosen based on the cone structure of the circuit; the scan chain is decomposed into several partitions, only one of which is controlled and observed at a time. To maximize the effectiveness of this scheme, we propose an ancillary test generation process wherein the detection of all the circuit faults is guaranteed within the shortest test time possible.

2 Previous Work

In recent years, a considerable amount of effort has been expended in reducing test application time and test data volume through test data compression [1, 2]. In both schemes, the test vectors are ordered so as to increase the correlation among these vectors and hence to obtain more compressed difference vectors through run-length [1] and Golomb [2] encoding schemes, respectively. The encoded difference vectors are decompressed through an on-chip circuitry; the actual test vectors, which are reconstructed on-chip, are then shifted into the scan chain. As the on-chip test data decompressor decouples the scan chain from the ATE, the test vectors can be shifted into the scan chain at a rate faster than the one supported by the ATE; however, the restrictions imposed on the clock rate due to test power considerations limit the applicability of these techniques. In [3], a test pattern compression methodology is proposed wherein a small number of virtual scan chains visible to the ATE are used to drive a large number of internal scan chains; loading the decompressed test data in parallel to a large number of scan chains significantly reduces test data volume. Even though significant test time and test data reductions are achieved, the large number of scan chains driven simultaneously results in excessive test power dissipation.

[*]The work of the first author has been supported through a graduate fellowship by IBM.

0-7803-7607-2/02/$17.00 ©2002 IEEE

The proposed solutions in the literature that aim at reducing test power have focused on reducing the switching activity in the circuit. The transitions that originate from the scan chain can be prevented from propagating into the circuit through the use of externally controlled gates [4]; however, such techniques result in performance degradation as they necessitate gate delay insertion on critical paths. In [5], scan chain modification through the insertion of logic gates between the scan cells is proposed to reduce the scan chain transitions associated with the test stimuli inserted. The actual test vectors are analyzed to identify the location and the type of logic to be inserted to the scan path. Even though considerable *scan-in* test power reductions are achieved with no performance penalty whatsoever, *scan-out* test power is not handled.

Several scan chain architectures have been proposed for test time, test data, and test power reduction. In [6], the scan chain is partitioned into several smaller ones. Activation of only one scan chain at a time restricts rippling, reducing the test power dissipated; however, such an architecture reduces neither test time nor test data. In [7], several scan chains are driven through a single test pin; only one of these chains receives deterministic test data whereas the remaining ones are fed from LFSRs in parallel, thus reducing test application time and test data volume. Similarly in [8], a single test pin is used to control several scan chain partitions. The majority of the faults are detected by loading all the partitions with the identical test stimulus; the remaining faults are detected by configuring the partitions as a single scan chain. In both [7] and [8], test time and test data reductions are achieved; however, test power considerations are overlooked.

3 Proposed scan architecture

At any point, the need to control scan cells is limited to the ones for which a specified value needs to be attained. The proposed scan-based testing scheme exploits this observation by decomposing the scan chain into several partitions, only one of which is controlled and observed at a time; the scan-in pin is directed to one of the partitions while simultaneously the content of this partition is observed through the scan-out pin. The test response data captured is frozen within the remaining partitions while the active partition is fully controlled and observed. The length of the test vector to be applied to the circuit under test may differ; depending on the scan cells that need to be controlled and observed, more than one partition may be loaded with deterministic data and observed while the remaining partitions inherit the preceding test response as the current test stimulus data. If more than a single partition needs to be controlled and observed, these partitions are directed to the scan-in (scan-out) pin in a sequential manner; only one scan chain partition is active at a time. The scan chain architecture is given in Figure 1. Prior to the control (observation) of any particular partition, the select lines of the DEMUX (MUX) should be set to the bits that identify the partition. These bits are conveyed through the TDI pin in a manner identical to the actual test data bits; the partition select bits contribute to the overall test application time as well.

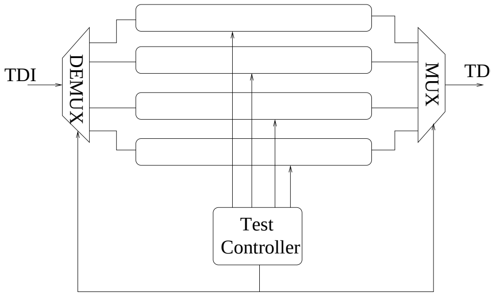

Figure 1. Proposed scan chain architecture

It should be noted that in traditional scan-based testing, all the scan cells are fully controlled and observed. Application of every test vector therefore necessitates the insertion of a number of bits that equals the number of scan cells. In the proposed scheme, however, every test vector necessitates control of a subset of scan cells, reducing the test time per test vector. The overall test time and test data reductions are determined by the length and the number of test vectors. The test time and test data, denoted as TAT and TDV, respectively, can be expressed as

$$TAT = TDV = \sum_{i=1}^{N} v_i(ilogN + mi + 1) \tag{1}$$

with m denoting the number of scan cells in a scan chain partition, v_i, the number of test stimuli that necessitate controlling i partitions, and N, the number of scan chain partitions. For each test stimulus vector, in addition to the shift-in cycles for the actual deterministic test data denoted by mi, and the capture cycle, several additional cycles are expended to convey the partition select bits.

Even though the proposed scheme achieves the insertion of the test stimulus data into the scan chain through the control of a subset of the scan cells, not all the fault effects activated by the stimulus vector are necessarily detected; only the ones that manifest in the partition directed to the scan-out pin are captured.[1] The proposed scheme can be extended through the utilization of a *spare capture register* so as to enable the detection of an increased number of faults per test stimulus. Figure 2 illustrates the scan chain architecture wherein an additional partition, the spare capture register, is utilized for capturing the *response signature*. The signature captured in the spare register consists of the response data in the partitions XORed bitwise. The bitwise XOR operation is piggybacked in the capture cycle; while the response data are captured in the scan chain partitions, the XOR of the response bits to be written into the same bit position of the partitions is captured in the corresponding spare register cell. The only partition that is observed throughout the whole test application process is the

[1] Additional checks should be performed to prevent the aliasing of the faults whose first manifestation occurs in an unobserved partition, as such faults result in the delivery of unintended stimuli.

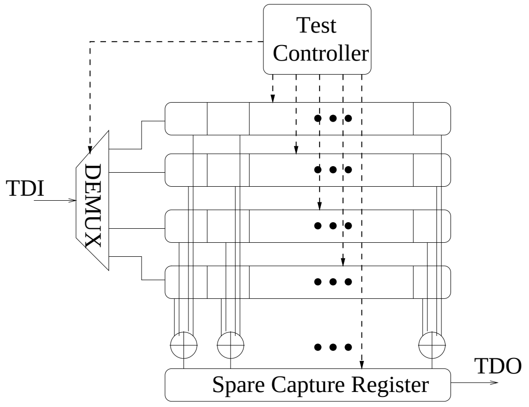

Figure 2. Spare register for response compression

spare capture register in this case to which the scan-out pin is connected.

Detection of a fault in the response signature in this scheme is assured only if an odd number of scan cells in the same bit position of the partitions display the effect of the fault. Non-detection of a fault necessitates its manifestation for every stimulus vector in an even number of scan cells for every bit position, a highly unlikely aliasing scenario.

The latter scan architecture, *i.e.*, the scheme with the spare register, imposes some area cost due to the XOR gates and the additional scan cells utilized for capturing the response signature. Compared to the initially proposed architecture, wherein several partitions remain unobserved, an increased number of faults are captured per stimulus, resulting in fewer test stimuli and hence shorter test time. The selection among the two versions of the methodology that we propose in this section is determined by the strictness of the area constraints; additional area cost for a spare register can be expended to further reduce test time and test data if additional space exists on the silicon die. A quantitative comparison of test time reductions and area costs associated with both methodologies is provided in the experimental data section.

4 Optimal scan chain partitioning

The effectiveness of the proposed scheme is determined by the distribution of scan cells to scan chain partitions. Application of test vectors that minimize the number of scan chain partitions to be controlled necessitates placement into the same partition of the scan cells that are likely to be simultaneously specified. The optimal decomposition of the scan chain depends on the cone structure of the particular circuit, consequently. To identify the scan cells that need to share the same partition, the associated test cubes can be analyzed; these test cubes implicitly represent the cone structure information.

A graph-based method is used to implement the optimal placement of the scan cells. The nodes denote scan cells and the edge weights represent the number of test cubes in which the corresponding cells are simultaneously justified to specified values. In every step of the algorithm, the scan cell pair with the maximal weighted edge connecting them is selected to be placed in the same scan chain partition.

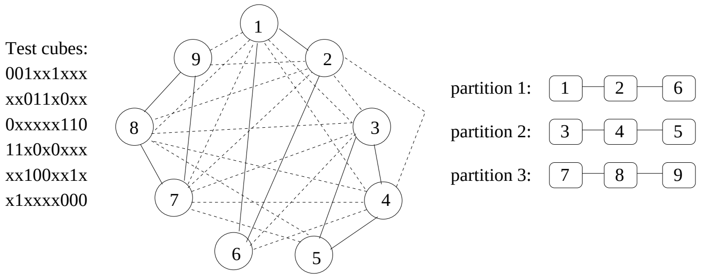

Figure 3. Scan chain partitioning

The example in Figure 3 illustrates the scan chain partitioning process. Based on the actual test cubes given, a graph is constructed that pinpoints the scan cell correlations; highly correlated scan cells are placed in the same partition as shown in the figure. In the graph, the dotted lines denote the edges with unit weight whereas the solid lines represent the ones with a weight of two; no pair of bits is specified in more than two cubes, in this example.

Even though placing the highly correlated scan cells in the same partition helps shorten the overall test time, layout constraints should be additionally considered. In obtaining the circuit layout from a high-level description, flip-flop-register correspondence information is utilized in stitching the scan cells; the layout construction is performed so as to minimize the wiring and routing overheads. These considerations can be incorporated into the scan chain partition problem through the placement of nearby scan cells in the same scan chain partition, resulting in the implementation of the near-optimal scan chain configuration with no layout constraint violations.

5 Test Generation

Detection of all the circuit faults by the fewest possible test vectors necessitates test generation that takes into account the scan chain configuration. As a test stimulus embeds parts of the previous test response in this scheme, test generation needs to incorporate values thus prespecified; the test generator is forced to justify the scan cells in the fewest possible partitions with the remaining inputs tied to the binary values implied by the previous test response.

Every time a test vector is generated, the partition to be controlled should be selected. The selection criteria consist of the number of faults detected by inserting deterministic data to the partition; this number cannot be computed straightforwardly as the number of faults detected by loading a scan chain partition with deterministic data depends on the data inserted. Instead, the number of faults that definitely remain undetected regardless of the inserted deterministic data can be calculated. For a given scan chain partition, the primary inputs in the remaining partitions are tied to the predecessor test response values and a straightforward logic simulation is performed; the faults with the unsatisfied activation or propagation requirements are thus identified. The example in Figure 4 illustrates two such faults when several primary inputs are tied to certain binary values. The stuck-at-0 fault on line b cannot be activated whereas the stuck-at-0 fault on line a cannot be propa-

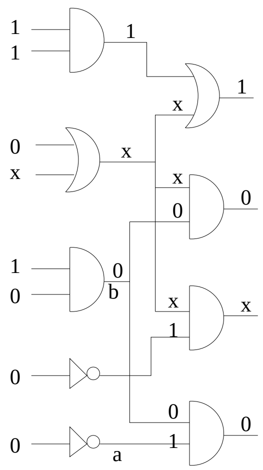

Figure 4. Identification of potentially detected faults

gated. Pinpointing the faults that definitely remain undetected implicitly helps identify *potentially detected* faults when the scan chain partition is loaded with the appropriate data.

Once the partition to be controlled is selected, the deterministic test data to be inserted to the partition should be computed. The deterministic test data is produced based on the primary input assignment for capturing the hardest-to-detect fault among the potentially detected faults corresponding to the scan chain partition. As the easier-to-detect faults frequently end up being caught in the process of testing harder faults, we prioritize the test generation process by starting from the hardest ones. The detection probabilities of the faults are computed based on the SCOAP [9] measures. The pseudocode for the test generation process is given in Figure 5.

6 Experimental Results

The proposed scan-test scheme has been implemented and applied to the fully-scanned circuits in ISCAS89 [10]. Both versions of the methodology we propose have been implemented; the associated scan architectures are given in Figures 1 and 2, respectively. The proposed test generation process is implemented by modifying the code of ATALANTA [11]. The experimental results in the case of the architecture with no spare capture register are given in table 1; as the area overhead is less than 0.5% for all the circuits, it is not explicitly reported. In the table, columns 2 through 5 provide the number of test vectors that require controlling one, two, three and four partitions, respectively. The number of test vectors generated by ATALANTA is provided in column 6 for test time and test power comparison purposes; the test application time reduction achieved by the proposed methodology is given in column 7. Columns 8 and 9 present the percentage reduction in total and average test power, respectively. The number of partitions that the scan chain is decomposed into is selected based on the number of bits that need to be controlled simultaneously for the application of test stimuli. Even though the optimal number of partitions may depend on circuit characteristics, we used four scan chain partitions as numerous experiments have shown that this configuration generally leads to the best results in terms of test time. Increasing the number of partitions reduces the length of each partition, resulting in fewer test cycles for controlling every partition; however, the number of unobserved scan cells that may possibly contain the fault effects is increased, degrading fault detection per test stimulus. As the test set generated by the proposed test process mostly consists of test stimuli that require only a subset of the partitions, significant test time reductions are achieved. For instance, for the circuit, $s13207$, the number of test stimuli that require controlling one, two, three and four partitions, is 413, 203, 11 and 1, respectively, resulting in a test application process that takes 144,878 cycles; on the other hand, application of a traditionally generated test set consisting of test stimuli that necessitate controlling every partition results in 307,071 test application cycles. Test power[2] reduction is achieved as well since only one out of four partitions is active at any time during test application; average test power reductions are around 75% as expected. As test time is reduced as well, total power reductions exceed average power reductions.

```
Test generation(Fault list)
While the fault list contains irredundant faults {
  For every partition
    Identify the potentially detected faults;
    Select the partition(s) to be accessed;
    Generate test under the constraint of all the uncontrolled inputs being tied to the previous test response values for the hardest-to-detect fault that is irredundant in this context;
    Perform fault simulation to drop from the fault list all the faults that manifest in the register(s) being observed
} return;
```

Figure 5. Test generation pseudocode

The second version of the proposed scheme, wherein a spare capture register is utilized, has also been implemented; the experimental results are provided in table 2. In contrast to the first version, an increased number of scan chain partitions may result in reduced overall test time; the spare register captures the effect of most of the activated faults even in the case of a fairly large number of partitions. Increasing the number of partitions not only only helps reduce the overall test time but furthermore the area cost is alleviated as the size of the spare register is decreased; however, the number of partition select bits, which also contribute to the overall test time, increases logarithmically with the number of partitions. In the second column of table 2, the number of partitions that leads to the shortest test time is given; the best configuration for every circuit is identified through numerous experiments. The reductions in test application time, total and average power are provided in columns 3 through 5 while the sixth column presents the area cost associated with the spare capture register and the XOR gates. Compared to the former version of the methodology we propose, the utilization of a spare register results in further reductions in test time; the number of faults detected per test stimulus is increased, decreasing test stimuli

[2] Test power is computed based on the number of scan chain transitions.

Circuit	v_1	v_2	v_3	v_4	T	TAT Red'n.(%)	Total Power Red'n.(%)	Avg. Power Red'n.(%)
s713	60	21	0	0	58	37.52	87.71	72.04
s953	110	0	0	0	92	61.86	93.53	78.37
s1423	22	41	25	9	64	6.94	75.15	70.41
s5378	107	106	85	55	252	17.71	73.20	65.98
s9234	122	187	118	95	367	13.27	80.08	76.22
s13207	413	203	11	1	459	52.82	90.95	80.59
s15850	344	244	52	4	442	42.45	84.83	73.29
s38417	754	573	107	0	882	36.74	84.32	75.09
s38584	597	349	67	4	653	41.79	86.83	77.25
Average	281.00	191.56	51.67	18.67	363.22	34.57	84.07	74.36

Table 1. Results of proposed scheme with no spare register

necessary. The larger number of partitions results in reduced rippling in the scan chain, leading to considerable reductions in test power. Practically, the test power problem disappears.

In the proposed scan testing scheme, no fault coverage degradation occurs whatsoever compared to the traditional scan test; even the faults whose detection necessitates controlling and observing all the scan cells are captured by the proposed scheme through the control and observation of every partition in a sequential manner.

7 Conclusion

Even though current scan test practices ease the testability problem through the conversion of sequential circuits to combinational ones, they impose high test time and test power costs. In this paper, we propose a new scan test scheme which delivers practical solutions to the aforementioned problems. A given scan chain is decomposed into partitions, only one of which is active at a time, while the remaining partitions are frozen with the response data captured. The high density of the unspecified bits in test stimuli enables the scheme we propose wherein a test stimulus partially consists of the previous test response data captured in the scan cells. Exploitation of the *don't care* bits in this manner delivers not only significant test power reductions, but test time improvements furthermore, as only a subset of the scan cells are controlled for the delivery of the test stimuli.

The effectiveness of the proposed scan test methodology is increased through the use of the test generation process customized for the scan architecture. Generation of a test stimulus based on the previous test response helps reduce the number of partitions that need to be controlled for the insertion of the test stimulus, shortening the test time. Several heuristics are embedded in the test generation process to appropriately select the partitions to control and the faults to handle.

		Reduction (%)			Area
Circuit	Part.	TAT	Tot. Power	Avg. Power	Cost(%)
s713	4	63.70	90.93	73.28	7.51
s953	4	78.82	92.21	76.81	8.20
s1423	4	69.43	92.36	77.01	9.22
s5378	8	77.93	97.24	87.51	7.21
s9234	8	82.33	97.79	88.92	5.17
s13207	16	92.25	99.70	92.41	7.98
s15850	16	90.82	99.61	94.38	6.60
s38417	32	93.49	99.86	96.71	7.16
s38584	32	94.76	99.89	97.03	6.71
Average	13.78	82.62	96.62	87.12	7.30

Table 2. Proposed scheme with spare register

The proposed scan-based testing methodology along with the customized test generation process have been implemented and applied to several fully-scanned circuits. The experimental results confirm the considerable test time, test data volume and test power reductions attained. Power-efficient, rapid test of circuits is realized by the scan architecture and the associated scan-based test scheme we propose.

References

[1] A. Jas and N. Touba, "Test Vector Decompression via Cyclical Scan Chains and Its Application to Testing Core-Based Designs", in *ITC*, pp. 458–464, 1998.

[2] A. Chandra and K. Chakrabarty, "System-on-a-chip Test-Data Compression Architectures Based on Golomb Codes", *IEEE TCAD*, vol. 20, n. 3, pp. 355–368, March 2001.

[3] I. Bayraktaroglu and A. Orailoglu, "Test Volume and Application Time Reduction Through Scan Chain Concealment", in *DAC*, pp. 151–155, 2001.

[4] H. J. Wunderlich and S. Gerstendorfer, "Minimized Power Consumption for Scan Based BIST", in *ITC*, pp. 85–94, 1999.

[5] O. Sinanoglu, I. Bayraktaroglu and A. Orailoglu, "Test Power Reduction Through Minimization of Scan Chain Transitions", in *VTS*, pp. 155–161, 2002.

[6] L. Whetsel, "Adapting Scan Architectures for Low Power Operation", in *ITC*, pp. 863–872, 2000.

[7] A. Jas, B. Pouya and N. Touba, "Virtual Scan Chains: A Means for Reducing Scan Length in Cores", in *VTS*, pp. 73–78, 2000.

[8] I. Hamzaoglu and J. H. Patel, "Reducing Test Application Time for Full Scan Embedded Cores", in *FTCS*, pp. 260–267, 1999.

[9] L. H. Goldstein, "Controllability/Observability Analysis of Digital Circuits", *IEEE TCAS*, vol. 26, n. 9, pp. 685–693, September 1979.

[10] F. Brglez, D. Bryan and K. Kozminski, "Combinational Profiles of Sequential Benchmark Circuits", *IEEE ISCAS*, vol. 14, n. 2, pp. 1929–1934, May 1989.

[11] H. K. Lee and D. S. Ha, *On the Generation of Test Patterns for Combinational Circuits*, Technical Report 12-93, Department of Electrical Eng., Virginia Polytechnic Institute and State University.