Low Power Illinois Scan Architecture for Simultaneous Power and Test Data Volume Reduction

Anshuman Chandra, Felix Ng and Rohit Kapur Synopsys, Inc., 700 E. Middlefield Rd., Mountain View, CA

Abstract

We present Low Power Illinois scan architecture (LPILS) to achieve power dissipation and test data volume reduction, simultaneously. By using the proposed scan architecture, dynamic power dissipation during scan testing in registers and combinational cells can be significantly reduced without modifying the clock tree of the design. The proposed architecture is independent of the ATPG patterns and imposes a very small combinational area penalty due to the logic added between the scan cells and the CUT. Experimental results for two industrial circuits show that we can simultaneously achieve up to 47% reduction in dynamic power dissipation due to switching and 10X test data volume reduction with LPILS over basic scan.

1. Introduction

As integrated circuit (IC) geometries shrink and gate oxide thickness of the transistor and operating voltage get smaller, controlling power dissipation has become very challenging. With ever increasing number of transistors on the IC, the test data required to test the circuits has also increased rapidly. Higher test data volume translates into higher test cost as it requires expensive testers with more memory to store the test patterns and longer test application time to apply the test patterns. The application of scan test patterns to the circuit causes a lot of switching activity in the flip flops of the scan chain and the CUT when the test patterns are shifted in and the test responses are shifted out. Both the scan shifting and the response capture power are known to be significant part of the power dissipation during test [1] [2] and that it exceeds the power dissipation in the functional mode.

A number of techniques to control power consumption in test mode have been presented in the literature. Structural methods that require design modifications for controlling power have been proposed in [2]-[6]. A decoder/multiplexer-based architecture for gating scan chains has been proposed in [2]. In [3], shift registers are used to gate portions of the scan chain during shifting while counters are used for gating in [4] to reduce average and peak power dissipation by transforming conventional scan architecture into desired number of selectable, separate scan paths. A solution for the average power

problem based on minimum transition fill vectors was proposed in [1] and for reducing peak power during external testing, a technique based on selective disabling of the scan chain was presented in [5]. In [8] the authors provide a solution to prevent peak power violation during both shift and capture cycle using scan chain partitioning.

A simple and effective solution to significantly reduce test power, independent of test set is to modify the scan cell and add gating logic to mask the scan path activity during shifting [7][8]. The blocking gates are controlled by the scan enable signal and the inputs to the CUT remain fixed at either logic 0 or logic 1 during the entire scan shift operation [7]. This approach coupled with random pattern suppression provides significant power savings during BIST [8]. The vector inhibiting technique presented in [8] provides a hardware solution to the power minimization problem and is shown to significantly decrease power consumption during BIST sessions.

ATPG techniques for generating vectors that lead to low power testing are described in [9] and [1]. However, while these techniques provide reduction in power consumption, they do not lead to any appreciable decrease in test data volume. Test generation for low-power scan testing usually leads to an increase in the number of test vectors [9]. On the other hand, static compaction of scan vectors causes significant increase in power consumption during testing [1]. A unified scheme for reducing test power, data volume and testing time based on test data compression was first presented in [10]. However, most of the scan power saving techniques proposed in the literature are either very design intrusive or are based on ATPG.

 Many techniques to reduce test data volume have been proposed in the literature and [11] provides a good survey on the different compression techniques. One such technique to address the test data volume problem is Illinois scan architecture (ILS) [12]; see Figure 1. ILS uses a single scan input to feed all the scan chains in the broadcast scan mode and the response is collected in a multiple input signature register (MISR) or an XOR compactor.

In this paper, we present the low power ILS (LPILS) scan architecture to address the problem of test data volume and test power, simultaneously. We enhance the broadcast scan mode and show that the compression and power reduction can be achieved together at a very nominal combinational area overhead. Since no sequential elements are added to the design, the proposed technique does not require clock tree re-synthesis. The rest of the paper is organized as follows: Section 2 describes the proposed LPILS architecture for reducing dynamic power dissipation in the scan chains and the combinational logic during scan shifting. Section 3 outlines the flow that we used to estimate power dissipation for different scan architectures using industry standard tools. Section 4 presents results for area, and power dissipation for two industrial circuits. Section 5 concludes the paper.

Figure 1: Block diagram of two configurations of Illinois scan.

2. Low Power ILS Architecture

We first briefly discuss the ILS scheme presented in [12] w.r.t. test data compression and power consumption during test. ILS architecture is very simple scan architecture for reducing test data volume. However, as demonstrated by the data presented in Section 3, the power dissipation during test in ILS, based on how the don't-care bits are filled, is almost the same or worse as compared to the basic scan architecture. This is because like basic scan, all the scan flops and the CUT are subjected to continues switching activity during shift and capture cycles in ILS. Due to similar switching activity profiles, the dynamic power dissipation for basic scan and ILS during test is also similar. ILS also suffers with the problem of fault coverage loss when high compression values are target. This limits ILS architectures test data volume reduction capabilities to be between 10X to 20X compression.

It is important to note that a unique property of broadcast scan mode of ILS is that at the end of a shift in cycle, the data in all the chains is exactly the same. Therefore, it does not matter how this data reaches the scan chains as long as we can ensure that before the capture cycle, all the chains contain the same scan data. This property of ILS is exploited in this paper to reduce dynamic power dissipation due to scan in switching activity.

We now describe the low power ILS architecture in detail. The serial scan mode of ILS is retained as it is. However, as shown in Figure 2, the broadcast scan mode is modified by dividing the scan chains into equal length blocks of size

b and adding a set of multiplexers to route the scan in data to the chains. Each scan block contains b scan cells. The first chain is referred to as the Reference chain is fed directly by the scan input SI and contains no multiplexers. All the remaining shared chains are either fed from the scan blocks in the same chain or the scan blocks in the Reference chain as shown in the Figure 2. The scan-in data flow is controlled by the $LPWR_{sel}$ signal, which selects the input for the scan block using the multiplexers preceding each scan block in the shared chains.

Figure 2: Block diagram for the broadcast scan mode of low power Illinois scan architecture.

For LPILS architecture with scan block size b and with n number of scan blocks per chain, the shift cycle length is equal to nb. During scan-in shift cycle, the Reference chain is used as a scan-in data cache for the first $(n-1)b$ shifts. For the first $(n-1)b$ shifts, the LPWR_{sel} signal is set to 1 causing the scan data on the SI signal to set values in the reference chain. While the reference chain is updated, in this configuration the shared chains receive constant data from the pseudo scan input that is connected to the ground and the first $n-1$ scan blocks in the shared chains are filled with value 0. While we were focused on the values shifting into the scan chains it should be noted that the captured values in the last $n-1$ blocks of each scan chain are observed simultaneously during the scan operation.

For the remaining b shifts, $LPWR_{sel}$ signal is set to 0. This routes the scan data in the scan blocks of the Reference chain to shift into the scan blocks in the shared chain through the multiplexers. The ith scan block in the shared chain receives data from the $(i-1)^{th}$ scan block in the Reference chain. The first scan block in the shared chains receives scan-in data from SI for the last b shifts. Note that the scan out data of each chain is shifted out normally for the entire shift cycle completing the observation of the remaining unobserved scan cells from the previous step. The test application proceeds as follows:

- 1. Set $LPWR_{sel} = 1$
- 2. Scan-in data into the Reference chain and the shared chains for $(n-1)b$ shifts while scanning out values from all the chains
- 3. Set $LPWR_{sel} = 0$
- 4. Scan-in data in the scan blocks of the shared chains from the previous scan block in the Reference chain for the last b shifts. For the first scan block, scan in data into Reference chain and shared chains from the SI.
- 5. Set the primary input values and the scan enable
- 6. Apply the capture clock to capture the response of the CUT
- 7. Repeat steps 1 through 6 for the next scan-in and scan-out cycle

We now discuss the scheme through a hypothetical example of three chains with 18 scan cells each and each chain is divided into 3 scan blocks of length 6 (see Figure 3). The scan blocks are colored in different colors to show the state of the chains at different times during the shift in operation. The orange color represents the response bits, blue color represents stimulus bits and the green color represents the pseudo scan input values. A lighter shade of blue is used for the scan input data that is shifted in while the $LPWR_{sel}$ signal is set to 1.

Figure 3: Block diagram of LPILS showing different stages of scan data shifting for $b = 6$ and $n = 3$. (a) Response captured in the scan flops. (b) State of scan chains after scan in shift for (n-1)b shifts. (c) State of scan chains after scan in shift for nb shifts.

Figure 3 shows the state of the scan chains at different stages of scan in shift operation for the broadcast scan mode. Figure 3 (a) shows the scan chains with the response of the previous pattern. Now the $LPWR_{sel}$ signal is set to 1 and the scan-in data is shifted in for $(n-1)b = 12$ cycles, where $n = 3$ and $b = 6$. As shown in Figure 3 (b), the scan in data is shifted into the Reference chain while the scan cells in the shared chains are filled with 0s. $LPWR_{sel}$ signal is set to 0 for the next 6 cycles and this causes the scan in data to fan out into the shared chains through the multiplexers; see Figure 3 (c).

As shown in Figure 3, the proposed scheme is guaranteed to reduce switching activity in the scan chain and the CUT

during scan in. The switching activity reduction achieved is independent of how the patterns are generated by ATPG. However, if the don't-care bits in the scan-in vector are mapped to specified bits using minimum transition count (MTC) algorithm [1][10], switching activity will be further reduced as compared to the random fill approach. The scheme does not require gating of the clocks or blocking of the scan cell output. The additional multiplexers added to the design increase the load on the last flop of each scan block of shared chains. But unlike the blocking logic techniques, the additional logic is not directly in the functional path and it is purely combinational. This is also an advantage over previous schemes [4], where comprehending complex sequential control logic in the clock path could be difficult for ATPG tools. Moreover, as shown by the experimental results, the additional hardware cost is very small.

The compression achieved by LPILS is the same as ILS. For example, feeding two shared chains as shown in Figure 3 usually translates into 3X compression. In general for ILS, feeding $(M-1)$ shared chains is expected to provide MX compression. However, when high compression values are targeted, the scan cell dependencies cause the compression to drop rapidly. Therefore, LPILS guarantees the best compression obtained by ILS without loss in coverage.

The other advantage of this scheme is that the scan chains need not be completely balanced. As long as the scan chains are divided in such a way that the first $(n-1)$ scan blocks are of equal length b , and the last scan block is of length b', where $b' > b$, the scheme would work properly. However, maximum switching activity reduction is obtained when $b' = b$. The LPWR_{sel} signal can be controlled in a similar way as the scan enable signal is controlled. A counter on the automatic test equipment (ATE) can be used to switch it from 0 to 1 at the end of $(n 1/b$ shift in cycles. Since $L PWR_{sel}$ is a slow changing signal, it can be even multiplexed with an existing slow speed functional pin.

3. Power Estimation Flow

In order to compute power reduction achieved using LPILS, we computed power dissipation numbers for basic scan, ILS and LPILS using industry standard tools. The flow shown in Figure 4 was used for estimating the power dissipation. The following tools were used in the flow shown in Figure 4:

- 1. We first synthesized DFT structures using Synopsys DFT Compiler;
- 2. We then generated test patterns using Synopsys TetraMAX® ATPG tool;
- 3. We then serially simulated the patterns using Synopsys VCS® simulator, capturing the activity of all the nodes of the design using VPD(VCD+) format. Serial simulation emulates the actual power dissipation when the test patterns are applied to

the CUT on the tester. No post-layout timing was back annotated during simulation.

4. We then used Synopsys PrimeTime® PX to analyze the power dissipation of the design. The inputs to PrimeTime® PX are: (i) CUT with DFT inserted; (ii) switching activity captured in VPD (VCD+) format in step 3 above; (iii) synthesis library containing power information for the chosen technology. PrimeTime PX can analyze the power dissipation by the types of cells in the design (combinational, sequential, clock tree) as well as the type of power dissipated (internal power, switching power and leakage power).

Figure 4: Flow used for power estimation.

The power analysis results obtained using this flow are far more accurate than using switching activity alone, because the characterized power dissipation information provided in the synthesis library are used for power analysis. The DFT synthesis and pattern generation for LPILS architecture required us to use special versions of DFT Compiler and TetraMAX ATPG tool. In addition, we manually verified the simulation waveform to ensure the $LPWR_{sel}$ signal is applied correctly and the patterns are shifted through scan chains properly; see Figure 5.

Figure 5: VCS simulation output with LPILS.

Figure 5 shows a part of the pattern simulation for three patterns. The first signal is $LPWR_{sel}$, next two signals are clock, fourth is a scan out pin, and remaining signals are the scan flops of the shared chain for which the scan out pin simulation data is shown. The flop closest to the scan in pin is at the top while the one closest to the scan out is at the bottom. As expected for the LPILS architecture, the flop closest to the scan-in pin is subjected to minimum switching activity whereas the flop closest to the scan out is subjected to maximum switching activity. The V shaped silent activity regions shown in the simulation are characteristic of LPILS broadcast scan mode and are responsible for reducing dynamic power dissipation in the shared scan chains.

4. Experimental Results

We implemented the flow on two industrial designs, CKT1 and CKT2. The netlists for the both designs were modified for adding the DFT logic for ILS and LPILS scheme. CKT1 has 199K gate primitives and 6.6K scan cells and CKT2 has 40K gate primitives and 4K scan cells, respectively. We used a 90 nanometer high performance library for our experiments. The power numbers reported in this section for total dynamic power also include clock tree power. Also, we did not observe any reduction in peak power, clock tree power or leakage power in our experiments.

4.1 Basic scan

To determine a baseline for calculating the reduction in power obtained using LPILS, we first of all generated data for basic scan with 10 scan chains using both random filling and MTC filling of Xs in the scan vectors for the two industrial circuits. Table 1 presents the number of scan patterns, test coverage, combinational, sequential and total dynamic power dissipation for both the circuits. It has been shown by previous research that MTC filling is very effective in reducing power dissipation due to switching activity in the scan chains and the CUT, and our results conform to that. We observe that MTC filling was able to reduce dynamic power dissipation during test by more than 50% for both the circuits. We also note that power dissipation in the sequential elements is far more than in the combinational logic. We use the total power dissipation using the random fill in Table 1 as the base line for calculating power reduction for the rest of the results.

4.2 Basic scan vs. ILS

Table 2 presents results comparing the basic scan dynamic power dissipation with ILS broadcast scan mode for both random and MTC filling with 10 scan chains. As shown in the random fill column of this table, ILS by itself results in virtually zero power savings. This is because random filling results in similar patterns for both basic scan and ILS and causes similar switching activity in all the scan chains and the CUT. The results for ILS and MTC filling are more interesting as we observe that the power savings obtained with ILS are far less than basic scan. For CKT1, ILS with MTC filling gives 30.59% reduction whereas basic scan with MTC filling gives 52.36% reduction in dynamic power dissipation. The reason for this difference is that broadcast scan mode is a far more constrained scan architecture, where setting a specified bit in one scan cell constrains all the cells in that column in each scan chain to the same value. Therefore, an ILS broadcast scan in pattern has fewer Xs as compared to basic scan pattern. Hence, the MTC filling does not result in significant reduction in dynamic power dissipation during test.

4.3 Basic scan vs. LPILS

We now discuss results for LPILS scheme presented in Table 3 and Table 4. Table 3 presents data where the target compression is fixed to 10X i.e., 10 scan chains are fed using a single scan input and the scan block size is varied from 5 to 20 in steps of 5. It is noted that 10X compression was obtained for both the circuits without significant loss in coverage. It is also observed that for both the circuits, more than 30% reduction was achieved using random fill and a suitable scan block size. The hardware overhead to achieve this reduction in power is also very small. In fact, for some cases, synthesis optimization was able to absorb the additional multiplexers into the existing logic and resulted in area savings. The important thing to note here is that 30% reduction in dynamic power was obtained with 10X compression and random fill vectors. This is significant because random filling of the vectors is very effective at catching nonmodeled defects. Therefore, LPILS provides the test engineer one way to address the problem of power dissipation, test data compression and catching nonmodeled defects, simultaneously.

Further power reduction can be obtained by using MTC filling. For example, for CKT2 with scan block size of 20, the increase in area is 1.74% and the reduction in dynamic power is 47%, which is almost 2X reduction in power due to switching activity. We also observe that as the scan block size is increased, power dissipation in both registers and combinational logic reduces irrespective of how the Xs are filled.

Table 4 presents data where scan block size is fixed to 10 and the target compression is varied from 5 to 20 in steps of 5. For random fill, we observe similar trend as in Table 3, where the scan block size was varied. As the target compression is increased, power reduction also increases. This is because the chains become shorter and require fewer shifts per test pattern. However, the scan cell dependencies in broadcast scan mode cause ILS to become inefficient at higher compression values and we observe a drop in fault coverage.

For MTC fill, as the target compression is increased, the power reduction obtained actually decreases; see Figure 6. Though, for a particular target compression, the dynamic power reduction is still more than random fill. This is very

interesting piece of data as it shows that for all the compression schemes based around shortening of the scan chains and loading them in parallel, MTC fill results in higher power dissipation as the number of chains are increased. This is again because lots of Xs in the scan in vector are now filled based on the constraints of the compression/decompression architecture and very few are left to take advantage of MTC fill.

Figure 6: Power reduction using LPILS and MTC fill.

5. Conclusions

In this paper, we presented LPILS scan architecture to achieve test data volume and test power reduction, simultaneously. LPILS scan architecture results in power savings in the broadcast scan mode independent of ATPG vectors, does not require clock tree resynthesis, and imposes a very small area penalty. We presented experimental data to show that 30% reduction in power due to switching and 10X compression can be obtained simultaneously with random fill. The reduction in power can be further improved to 47% if MTC fill is used. We showed that ILS by itself does not result in lower power dissipation during test. Finally, we also presented data to show that as higher compression is targeted, scan chains become shorter and the power savings obtained by MTC fill also decreases gradually.

6. References

- [1] R. Sankaralingam, R. R. Oruganti, and N. A. Touba, "Static compaction techniques to control scan vector power dissipation," in Proc. VLSI Test Symp., 2000.
- [2] J. Saxena, K. Butler, and L. Whetsel, "An analysis of power reduction techniques in scan testing," in Proc. Int. Test Conf., 2001, pp. 670–677.
- [3] N. Nicolici and B. M. Al-Hashimi, "Scan latch partitioning into multiple scan chains for power minimization in full scan sequential circuits," in Proc. of IEEE/ACM Design, Automation & Test in Europe Conf. , pp. 715–722, Mar. 2000.
- [4] L. Whetsel, "Adapting scan architectures for low power operation," in Proc. of Int. Test Conf. , pp. 863–872, 2000.
- [5] R. Sankaralingam et al., Reducing Power Dissipation During Test Using Scan Chain Disable, Proc. of VLSI Test Symp. , pp. 319–324, 2001.
- [6] P. M. Rosinger, B. M. Al-Hashimi and N. Nicolici, "Scan architecture for shift and capture cycle power reductions,"

Proc. Int. Symp. on Defect and Fault Tolerance in VLSI Sys., pp. 129–137, 2002.

- [7] S. Gerstendörfer and H. -J. Wunderlich, "Minimized power consumption for scan-based BIST," in Proc. of Int. Test Conf., 1999, pp. 77–84.
- [8] P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, "A test vector inhibiting technique for low energy BIST design," Proc. of VLSI Test Symp., 1999, pp. 407–412.
- [9] S. Wang and S. K. Gupta, "ATPG for heat dissipation minimization during scan testing," in Proc. ACM/IEEE Design Automation Conf., 1997, pp. 614–619.
- [10] A. Chandra and K. Chakrabarty, "A unified approach to reduce SOC test data volume, scan power and testing time," IEEE Trans. on Computer-Aided Design, vol. 22, pp. 352– 362, March 2003.
- [11] N. A. Touba, "Survey of test vector compression techniques," IEEE Design & Test of Computers, vol. 23, pp. 294–303, July-Aug. 2006.
- [12] A. R. Pandey and J. H. Patel, "An incremental algorithm for test generation in Illinois scan architecture based designs," Proc. of Design, Automation & Test in Europe, pp. 368– 375, 2002.

				Random fill				MTC fill				
					Comb.	Seq.	Total		Comb.	Seq.	Total	Power
	Scan	Gate	Cover-		Power	Power	Power		Power	Power	Power	Reduc-
	Elements	Primitives	age $(\%)$	Pat	(W)	(W)	(W)	Pat	(W)	W	(W)	tion $(\%)$
Ckt 1	6600	199000	100	143	8.06E-4	$1.71E-3$	$2.52E-3$	138	.57E-4	1.04E-3	$.20E-3$	52.36
						(31.97%) (68.03%)	100%			(13.07%) (86.93%)	(100%)	
Ckt ₂	4200	40000	99.95 234		2.84E-4	1.31E-3	59E-3	237	5.35E-5	$6.72E - 4$	$7.25E-4$	54.37
						(17.86%) (82.14%)	(100%)			(7.37%) (92.63%)	100%	

Table 1: Dynamic power dissipation for basic scan with random and MTC filling.

Table 3: Dynamic power dissipation for LPILS with varying scan block size and 10 scan chains.

Table 4: Dynamic power dissipation for LPILS with varying scan chains and scan block size = 10.

