A Critical-Path-Aware Partial Gating Approach for Test Power Reduction

MOHAMMED ELSHOUKRY
University of Maryland
MOHAMMAD TEHRANIPOOR
University of Connecticut
and
C. P. RAVIKUMAR
Texas Instruments India

Power reduction during test application is important from the viewpoint of chip reliability and for obtaining correct test results. One of the ways to reduce scan test power is to block transitions propagating from the outputs of scan cells through combinational logic. In order to accomplish this, some researchers have proposed setting primary inputs to appropriate values or adding extra gates at the outputs of scan cells. In this article, we point out the limitations of such *full gating* techniques in terms of area overhead and performance degradation. We propose an alternate solution where a partial set of scan cells is gated. A subset of scan cells is selected to give maximum reduction in test power within a given area constraint. An alternate formulation of the problem is to treat maximum permitted test power as a constraint and achieve a test power that is within this limit using the fewest number of gated scan cells, thereby leading to the least impact in area overhead. Our problem formulation also comprehends performance constraints and prevents the inclusion of gating points on critical paths. The area overhead is predictable and closely corresponds to the average power reduction.

Categories and Subject Descriptors: B.8.1 [**Performance and Reliability**]: Reliability, Testing, and Fault-Tolerance

General Terms: Algorithms, Design, Economics, Experimentation, Performance, Reliability

Additional Key Words and Phrases: Low-power testing, scan testing, scan cell gating, partial gating

1. INTRODUCTION

Power consumption during testing has become an important issue in modern-day designs. It is increasingly higher than power during normal functional operation [Zorian 1993; Wang and Gupta 1994]. In combinational circuits, consecutively applied test patterns have low correlation between them, and many of the input transitions may represent invalid or unlikely transitions during functional operation. In scan-based designs, many of the states that occur during shifting do not represent valid states in functional mode, and many of the state transitions in scan mode may never arise during normal operation. In test mode, test patterns are targeted towards faults and fault models with disregard to the circuit function, and pattern generation schemes try to exercise as many of these faults at once as possible to reduce the number of patterns, test time, and therefore cost. In functional mode, the principle of locality usually holds and only a certain percentage of the chip is active at-a-time.

Power consumption is especially important in today's chips, where larger numbers of transistors are packed into a smaller die size, higher frequencies are used, and aggressive timing requirements make at-speed testing methods highly important and frequently used [Girard 2002].

The abnormal power consumption during test can lead to adverse effects on the chip and the testing process, as outlined in Zorian [1993], Wang and Gupta [1994], and Girard [2002]. These include:

(1) Possibility of chip destruction due to excessive heat and the absence of a proper mechanism to dissipate this heat. Expensive packaging requirements or special cooling equipment are required to prevent this.
(2) Reliability problems due to high currents and elevated temperatures can accelerate destructive phenomena such as electromigration and lead to long- or short-term malfunction.
(3) Noise problems such as IR and Ldi/dt drops can cause the chip to falsely fail the test, a potential source of yield loss.
(4) Battery life in portable or remotely installed devices that need periodic testing.
(5) Since the package plays an important role in dissipating heat, high power consumption makes it difficult to obtain a carefully tested bare die to be used in multichip modules (MCM) or what is called the Known Good Die problem (KGD).

For all these reasons, various techniques have been proposed to reduce the impact of high power consumption during test application. The simplest of these are ad hoc techniques such as slowing down the speed of the test clock, partitioning the circuit into blocks that are separately tested in serial fashion, or

providing extra packaging and cooling. Since these will have negative impacts on test time and cost, and may not solve problems related to peak power, other methods were investigated for which we provide a brief overview. A good survey of many of these techniques can be found in Girard [2002].

1.1 Prior Work

Test scheduling algorithms have been proposed in Zorian [1993], Chou et al. [1997], and Ravikumar et al. [2000] to select a maximal set of tests that can be performed simultaneously under given power constraints. An ATPG algorithm has been suggested in Wang and Gupta [1994] for testing combinational circuits, and in Wang and Gupta [1997] for scan-based testing. Test pattern ordering has been discussed in Chakravarty and Dabholkar [1994] and Girard et al. [1999b, 1998], and scan cell ordering has been proposed in Chakravarty and Dabholkar [1994]. A compaction-based approach was introduced in Sankaralingam et al. [2000], where the order of merging test cubes is carefully selected to reduce power. Techniques in which the scan chain is modified to disable parts of itself while shifting in others are presented in Sankaralingam et al. [2001] and Whetsel [2000]. Scan chain partitioning into multiple scan chains is discussed in Nicolici and Al-Hashimi [2000], Saxena et al. [2001], and Ghosh et al. [2003]. Circuit partitioning for BIST designs has been discussed in Girard et al. [2000, 1999a], and a modified low-power LFSR for BIST-based applications has been shown in Ahmed et al. [2004].

For scan-based designs in particular, switching activity in the combinational part contributes to a large portion of the total switching activity in the circuit [Gerstendorfer and Wunderlich 1999]. These transitions are redundant during scan mode and are worth suppressing. A number of test power reduction techniques aim at reducing the number of transitions in the combinational part by blocking transitions occurring at scan cell outputs during scan mode. In Wang and Gupta [1997] and Huang and Lee [2001], input control techniques have been suggested in which a certain assignment for primary inputs is selected in such a way as to block transitions at as many gates as possible. In Huang and Lee [2001], an algorithm similar to the D-Algorithm [Bushnell and Agrawal 2000] has been used, where circuit nodes are ordered by their fanout and then a justification procedure is carried out to obtain a primary input assignment that maximizes the number of gates that do not switch, namely, those gates that have their controlling value determined by primary inputs. In Wang and Gupta [1997], Kernighan-Lin (K-L) iterative improvement has been used to attempt optimizing the almost always conflicting assignments of primary inputs that achieve the same objective.

A number of other techniques try to minimize transitions in the combinational part by gating the outputs of scan cells. A modified scan element has been suggested in Gerstendorfer and Wunderlich [1999]. By providing an extra gate at the output of each scan element, the output is held at a constant value during scan-in, and transitions in the scan flip-flops do not propagate to combinational logic. The gate is transparent in capture mode and during normal operation. In Zhang and Roy [2000], multiplexers have been used as gating elements, and

in Bhunia et al. [2004], a supply gating transistor has been introduced so as to turn off the first level of logic connected to scan cell outputs while in scan mode. In the last method, the area overhead is much less compared to adding an extra gate, as the area of a transistor is small compared to a full gate, and a single transistor can be shared by many logic gates. Moreover, this is required only for the first-level logic. However, the technique poses challenges to physical design and timing closure, and in the case where the transistor is sized to reduce delay, the area and delay overhead increase.

1.2 Contribution and Article Organization

In this article, we propose a new partial gating technique where only a subset of scan cells are gated. We show that with the proper selection of: (1) the percentage of gating elements, (2) their position in the scan chain, and (3) the output values at which to hold these gating elements, we can control scan power reduction to any desired level. In this partial gating approach, as opposed to full gating approaches, we do not sacrifice as much area, and we can minimize the impact on delay. Often, maximum power reduction is not required in favor of area and performance, and also for the purpose of keeping the circuit under test in conditions similar to its operating conditions [Butler et al. 2004]. We show that for almost the same number of gating elements, the maximum achieved average power reduction can be more than twice the minimum reduction, for the same area overhead, only by varying the locations and output of the gating elements. We describe an efficient way to evaluate how good a certain gating element placement is, and how closely this measure reflects the percentage of power reduction that can be achieved. We propose techniques to reduce the increase in peak power, a side effect of gating. Critical path information is used to avoid placing gating elements on critical paths. Our proposed method is not computationally intensive and the area overhead is predictable since power reduction and area overhead are closely related, as will be shown. We also point out the limitations of techniques that rely on controlling primary inputs only, and highlight the disadvantages of full gating and gating in general.

The article is organized as follows. Section 2 provides the background and a general introduction to our approach. Section 3 describes our proposed method for partial gating. Section 4 presents experimental results of implementing partial gating on several benchmark circuits. Section 5 presents our proposed techniques for peak power handling and associated experimental results. We provide our conclusions in Section 6.

2. BACKGROUND

In full scan testing, the power dissipated in the circuit can be divided into two components, namely, that dissipated in the scan flip-flops and that dissipated in combinational logic. These, in turn, can be categorized into different components [Dabholkar et al. 1998; Sankaralingam and Touba 2002a]: (i) power consumed during scan shifting, and (ii) power consumed during scan capture. Other sources of power consumption include power consumed by the clock network, interconnects, and leakage power. In CMOS gates, the main source of

power consumption is due to transitions at the output. The power consumed then is mainly due to the charging and discharging of parasitic capacitances, including the self-capacitance of the gate, the input capacitance of its fanout gates, and wiring. Also, depending on the slew of inputs, there is a time-frame when both the PMOS and NMOS subcircuits are both turned on, creating a low-resistance path from VDD to ground. The current that flows during this period is called short-circuit current. As logic gates get faster, short-circuit current contribution becomes less, and the main source of switching power is the charging and discharging of parasitic capacitances. Since the energy consumed in charging or discharing a capacitance C to/from a voltage V is $\frac{1}{2}CV^2$, the power dissipated is directly proportional to the output capacitance of the gate. The various methods proposed in the literature for estimating power using weighted transitions [Gerstendorfer and Wunderlich 1999] assume that gate capacitances are equal and ignore the wiring capacitance. We shall also follow this approach, while noting the limitations of relying exclusively on this approach as a measure of power. We also assume that transitions are fast enough to ignore the contribution of short-circuit current.

In order to reduce or eliminate transitions in combinational logic, transitions should be prevented from propagating as close as possible to the outputs of scan elements. In the next subsection we discuss gate controllability and how it can be applied for the purpose of reducing power, then in Section 2.2 we show the relation between gating selection and circuit transitions as a motivation for partial gating.

2.1 Gate Controllability

To prevent transitions at the scan cell from propagating to circuit logic, some degree of controllability is needed at the gates which receive their input from scan cells. In Wang and Gupta [1997] and Huang and Lee [2001], the authors proposed controlling primary inputs (PI) towards this purpose.

Several problems exist with approaches that try to control circuit test power using only primary inputs. First, in order to test the circuit properly, certain patterns have to be applied to primary inputs for each scan pattern. These patterns are generated by the ATPG tool and are necessary for proper fault detection. The generation of a certain pattern to be applied to PI during scan-in means that for each full scan, primary inputs have to change from the ATPG pattern to the low-power pattern after each capture cycle and before the next pattern is scanned. As was observed in our experiments, and also in Huang and Lee [2001], this extra switching event may actually lead to an increase in both average and peak power in some circuits. Moreover, the improvement gained by this low-power pattern over an ATPG pattern is not always significant and hence can do more harm by injecting the extra switching.

Another problem is that in large circuits, only a small percentage of gates can be controlled by PI, either directly or indirectly. Table I shows some statistics for a number of ISCAS'89 benchmark circuits to illustrate this point. Columns 2, 3, and 4, respectively, show the number of PI, flip-flops, and logic gates in the benchmark circuits. Column 5 shows the ratio of the number of PI to the flip-flop

Table I. Input Statistics for Some ISCAS'89 Benchmark Circuits

B.M.	#P	#FFs	#Gates	PI/FFs	%PI/Gates	%(PI+FF) Gates	Avg	Max	Improv. diff.	%diff
s5378	35	179	2779	20%	1.26%	7.70%	578	617	39	1.40%
s9234	36	211	5597	17%	0.64%	4.41%	1157	1801	644	11.51%
s13207	62	638	7951	10%	0.78%	8.80%	1295	1470	175	2.20%
s15850	77	534	9772	14%	0.79%	6.25%	2973	4345	1372	14.04%
s35932	35	1728	16065	2%	0.22%	10.97%	3631	4908	1277	7.95%
s38417	28	1636	22179	1%	0.13%	7.50%	708	1006	298	1.34%
s38584	38	1426	19253	3%	0.20%	7.60%	5100	5887	787	4.09%

count (FF), and we observe that this ratio is small for large circuits. Column 6 shows the ratio of PI to the gate count, and Column 7 shows the ratio of (PI+FF) to the gate count. It is clear that PI ratio is low, indicating that the controllability through PI is small. Columns 8 and 9 show the average and maximum number of PI-controlled gates, respectively. These numbers are only indicative figures, since they are obtained through simulations using randomly generated patterns for PI assignment. The last two columns show that the improvement of the best result (maximum number of PI-controlled gates) over the average is small. The average value corresponds to a typical ATPG pattern, whereas the maximum value corresponds to a case where the PI are set to a low-power vector during scan mode, as in Huang and Lee [2001]. When the difference between the maximum and average number of PI-controlled gates is small, and also small compared to the total number of gates in the circuit, then we can expect minor reduction in the average power through PI control. Although the average power is reduced, there may be an increase in peak power if the pattern switching causes a large number of signals to change at the same time.

A third problem is that such a method will require the PI to be set to some binary pattern. In low-cost testers, the number of interface pins is small, and it is customary to place boundary scan cells and make the PI also part of scan chains [Synopsys 2004b]. For large circuits where the number of PI is much smaller than the number of flip-flops (Table I), the area overhead in making the PI part of the scan chains is small, and doing so is a practical solution. Yet this will reduce the controllability for the PI, and the only way we can set PI to known values is through serial shifting. This problem can be overcome by placing multiplexing logic to route scan data, either to the boundary scan cells or internal scan chains. The advantage of the technique is that when the PI are being programmed, the scan cells are in fixed states and vice versa, easing power consumption. The drawback of the solution is that it incurs area and performance overhead, and adds complexities in the programming of the tester.

2.2 Gating Points and Circuit Transitions

As long as all the gates at the first level of the combinational logic are controlled (full gating), the actual values at the outputs of the gates are immaterial from a power perspective. However, if only a partial subset of these gates is controlled to minimize overheads (partial gating), the actual values on outputs of the controlled gates become important. Adding gating elements results in the
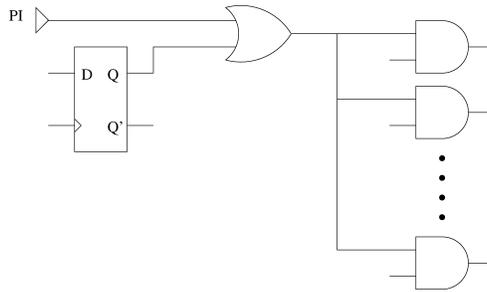
Fig. 1.   AND gates at the fanout of an OR gate.

following overheads: (a) Propagation delay in the gating element may change critical path delays in the circuit; (b) gating elements result in area overhead, which can be significant if the number of scan cells is large; and (c) peak power dissipation in the circuit goes up as the gating elements change between blocking and transparent modes. The selection of a subset of flip-flops for gating must be done with the aim of minimizing these overheads. In Sankaralingam and Touba [2002b], gating was performed with the goal of eliminating peak power violations. The method, however, requires extensive simulation of the circuit for all patterns to identify vectors that violate the peak power constraint, and then pattern simulation for each possible setting of control points, which makes it computationally intensive. It does not address average power, and the area overhead is unpredictable.

Inhibiting the transitions at the gate immediately following a scan flip-flop f does not guarantee that all transitions in the fanout cone of f will also be suppressed when the partial gating strategy is chosen. Consider the example of Figure 1, where an OR gate is used at the output of the flip-flop, and the output of the OR gate is controlled to logic 1 by setting PI = 1. The OR gate feeds a large number of AND gates, and potentially all of them can toggle if the other inputs feeding the AND gates toggle during scan shift. On the other hand, if the output of the OR gate is controlled to 0 by setting PI = 0 and Q = 0, the toggling on the outputs of AND gates can be prevented. Huang and Lee [2001] used a similar justification procedure for controlling power dissipation during scan test. However, their technique is to use only the PI for controllability, and we have explained earlier why this is not effective in large circuits.

In this article, we propose the use of a partial set of scan flip-flop outputs, which may be viewed as *pseudoinputs*, to control toggling activity in the combinational portion of the circuit. While the proposed partial gating method offers considerable flexibility, it also vastly increases the search space of possible solutions. In a circuit with p primary inputs and n scan flip-flops, it is easy to see that there are 2^{p+n} solutions when both PI and flip-flops are controllable, as opposed to 2^p solutions, when only the PI are controllable ($p \ll n$ in modern designs, see Table I). The assignment that leads to lowest power is dependent on the circuit structure and must be solved through optimization techniques. In the next section we will develop procedures to perform this.

3. GATING POINT SELECTION AND OPTIMIZATION

As explained in the previous section, reaching an assignment of binary values to the PI and pseudoinputs to minimize toggling activity in the combinational circuit is a difficult combinatorial optimization problem, even when some of the pseudoinputs are not gated. A cost function is needed to compare one assignment against another. In the next subsection we discuss our proposed cost function and in the following subsections we describe how to select and control the gating assignment.

3.1 Cost Function

An exact measure will involve counting the number of toggles in the combinational circuit, but the computational complexity of implementing such a metric is high, since it will involve detailed logic simulation of the circuit for every pattern and every assignment.

Two important points arise: One is that we need a cost function that is pattern-independent, since test patterns will continuously change for each scan/capture cycle and for different fault models. Second, it should not require extensive runtimes and simulation to evaluate a gating elements assignment. This gives motivation to our proposed method. We will take advantage of the circuit structure to deduce certain properties. First, our cost function will depend on the number of gates that will not switch during scan-in, with the assumption that scan cell outputs are continuously changing. In other words, those gates which are guaranteed to have a fixed output under our particular gating configuration that we are evaluating, and regardless of the contents of the scan chains, are considered. In such a case, the assignment we are evaluating will set some of the logic gate inputs to certain controlling values that will keep their outputs fixed, even if the other inputs were fed directly or indirectly from scan flip-flops and were changing with each shift.

Counting the number of these *controlled* gates, however, is not the best possibility, since it will falsely treat all gate transitions equally. In reality, transitions at gates outputs vary in their effect on power. A gate with high fanout will consume more power in a transition than one with a lower fanout since the load capacitance that is charged or discharged during the transition is larger. Consequently, we need to incorporate knowledge of fanout into our cost function so that assignments which are known to block more of those high fanout gates will give a higher merit.

Given the previous discussion, we conclude that we need to use a measure that attaches a weight to the number of gates guaranteed *not to switch* during the process of scanning-in a pattern. The weight attached to a gate g that satisfies this criterion is the fanout of g. In other words, our cost function for a pattern is of the form

$$Cost = \sum_{g \mid output(g)=1 \; \| \; output(g)=0} fanout(g). \tag{1}$$

Ideally, the weight can include the exact capacitance driven by the output of g. The capacitance information is not available until later in the design flow, after

---

COMPUTE_COST_FUNCTION (*V*)

/* V is a scan pattern that assigns 0, 1, or don't-care values *X* to PI and pseudoinputs.
An *X* at a pseudoinput corresponds to a nongated scan flip-flop.
It is assumed that *X* will lead to a toggling at the output of the scan flip-flop. */

    (1)   *cost* = 0

    (2)   Perform a three-valued logic simulation for vector *V*

    (3)   **for each** gate *g* **do**

      (a)   **if** (output (*g*) = 1 **or** output(*g*) = 0) **then** *cost* = *cost* + fanout(*g*)

    (4)   **return** *cost*

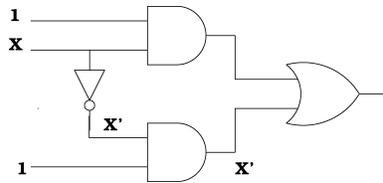Fig. 2.   Procedure to compute cost function.

Fig. 3.   Fixed output gate without a controlling input.

all the interconnects have been routed, and the fanout of the gate output can be taken as a measure of this capacitance in the absence of this information. A simple static technique to compute the cost function is shown in Figure 2. The cost function is a measure of the power savings resulting from partial gating of the scan flip-flops for a scan vector V.

We start with a vector V representing the PI and pseudoinputs. We are evaluating this vector as a possible gating assignment so as to compare it with other vectors representing different assignments. Vector V will have some bits as zeros, ones, or don't-cares. If a bit contains a one or zero, the corresponding input will be fixed at this value during scan-in. If it is a pseudoinput, a gate will be placed at the output of the corresponding flip-flop. If a bit contains a don't-care (X), it means that the corresponding scan flip-flop output will be free to change and no gate will be placed in front. Don't-cares will appear only if partial gating is used, and hence some of the scan flip-flops will have gating elements in front of them and some not.

After the vector V is specified, as we will see in the next subsection, the evaluation process starts with three-valued logic simulation of this vector. After the simulation, all gates that are fixed at 1 or 0 are summed, weighted by their fanout. These gates are not affected by transitions occurring at the output of nongated scan flip-flops. The three-valued logic simulation used in cost estimation has its limitations, although it is fast. For example, in Figure 3, the simulation will yield an X at the output of the inverter, leading to X at the output of AND gates, and an X at the output of the OR gate. Using a logic value such as $\bar{X}$ in the simulation will correctly predict the output at the OR gate as $X + \bar{X} = 1$. Thus, the procedure may underestimate the power savings resulting from partial gating. Similarly, the procedure does not perform
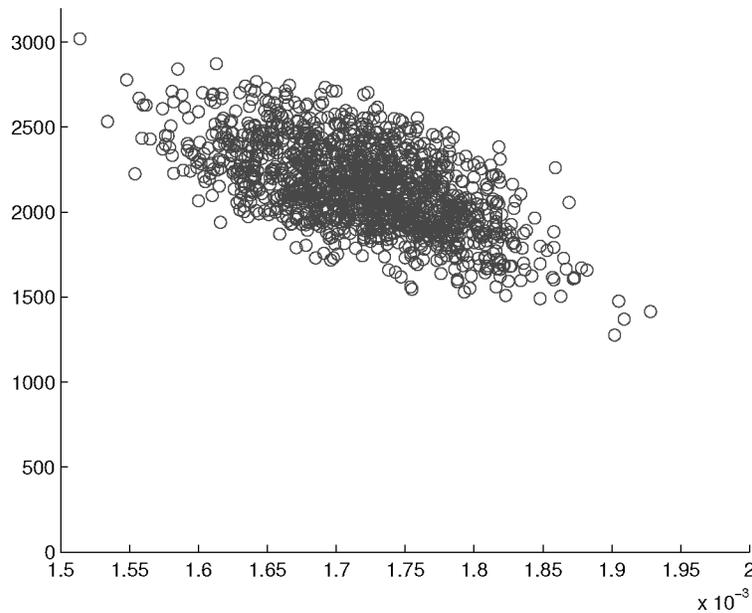
Fig. 4.   Cost versus power.

a delay simulation of the gates and hence cannot predict any glitching. In the example of Figure 3, the delay in the inverter can lead to a glitch at the output of the OR gate. Ignoring glitching can result in overestimation of power savings. Because of these two opposite effects, the estimate of cost function can be expected to be reasonably accurate, especially when used in comparing different gating assignments rather than obtaining a certain figure for power consumption.

To show the effectiveness of the cost function, we applied a large number of different gating assignments to a sample circuit (s5378), measured the cost function for each assignment, and then the average power due to this assignment. Figure 4 shows a scatter plot of the result. It is clear that there's a trend of correspondence between the cost function and the power with some perturbations. On the two extremes of the plot we see that the minimum cost function corresponds to maximum power among the evaluated set, and maximum cost function corresponds to the minimum power. This can only occur by applying a large number of different assignments, as required by our algorithm which is shown in the next subsection. Very low cost function values will mostly correspond to high power and vice versa, with some exceptions. These exceptions can be avoided by taking a bigger sample rather than just a single BestVector, and comparing them.

In the worst case, if none of the flip-flops are gated, the cost function for a vector can be zero. To completely eliminate toggling during scan shift, we can gate all the scan flip-flops. Between these two extreme cases are intermediate optimal solutions which keep the scan test power at a level that can be tolerated and close to its normal operation conditions [Butler et al. 2004].

INPUT_ASSIGNMENT

    (1)   Remove the scan flip-flops and convert them into pseudoinputs

    (2)   Read Critical Path information and compute the probability vector *probX*

    (3)   *MaxResult* $= -1$;

    (4)   *MinResult* $= \infty$;

    (5)   **for** $k = 1$ **to** *Number of Iterations* **begin**

     (a)   Generate a random vector $V[1 : p + n]$ of $p + n$ bits

     (b)   **for** $i = 1$ **to** $p + n$ **do**

       i.   Set $V[i]$ to X with probability $probX_i$;

     (c)   *Result* $=$ COMPUTE_COST_FUNCTION(V);

     (d)   **if** *Result* $>$ *MaxResult* **then** *BestVector* $= V$;

     (e)   **if** *Result* $<$ *MaxResult* **then** *WorstVector* $= V$;

    (6)   **end**

    (7)   **return** *BestVector*, *WorstVector*

Fig. 5.   Input assignment procedure.

3.2 PI and Gating Assignment

Formally, the input assignment problem is defined as follows. Given a full scan circuit with p primary inputs, n scan flip-flops, and a certain area overhead given as a certain percentage of scan cells, we wish to identify the subset of flip-flops whose outputs must be gated such that the savings in scan test power is maximum within the given area constraint. We also need to identify appropriate output values for these gating elements. An assignment procedure based on random search is shown next. The procedure repeatedly generates random vectors of size $p + n$ and samples a large population of the total search space of 2^{p+n} vectors. The bit $V[i]$ is set to an X with a probability $probX_i$ which is generally equal to $1 - (area\ overhead\ percentage)$. In the next subsection we will show how to estimate $probX$ if the given constraint is the required power reduction.

The generated vector is evaluated for its power-savings metric using the cost function of the previous subsection. A lower value for probability $probX_i$ implies a higher probability of inserting a gating element at scan flip-flop i. In this way we can control the area overhead, since on average there will be approximately $(1 - probX)*100\%$ gated flip-flops. The procedure can be easily modified to add an exact number of gated flip-flops with no major effect on the results. By keeping a limit in the loop 5-b (see Figure 5) which, if reached, the loop is exited, or if not reached, after exiting the loop, extra gating elements are added at random.

In this procedure we see that we aim to obtain the best vector *BestVector* which will give us best power savings among all assignments under the same area constraint specified by *probX*. *WorstVector* is only obtained for comparison purposes so as to demonstrate the correspondence between cost function and actual power consumed, as we will see in Section 4. Notice that *BestVector* and *WorstVector* are only meaningful in the context of partial gating where there is a large number of possibilities for where to place the gating elements and which values at which they should be held. In full gating, however, all scan cells are gated and hence combinational logic is totally turned off during scan-in, regardless of what values exist at the gate outputs, and hence there is no maximum or minimum cost.

3.3 Computing *probX*

$probX_i$ can either be set as a single value for all inputs or controlled independently for each input, forming a probability vector. In this way we can control the placement of gating elements at certain points. For example, for scan elements on the critical paths we can assign don't-care probabilities of one. In such a case, the algorithm generating bit patterns will always place an X at this bit position and never adds a gating element at the corresponding scan cell. Critical path information can obtained from timing analysis tools. For example, for the last three benchmarks shown in Table I, we obtained critical path information within a 5% window of the maximum delay. Using this information, the algorithm always places a don't-care at the output of each of these flip-flops on the critical paths. The observed increase in power was less than 2% of the power obtained without using critical path information, even on the benchamark circuit s35932 that had 32 flip-flops on critical paths. Also the number of critical paths for these circuits did not change after placing the gating elements.

The general value of $probX_i$ can be estimated from the percentage of average power reduction required in the combinational part. As will be shown in Section 4, for low to medium values of the power constraint, the area overhead is approximately around this value with around 10% variation, depending on circuit complexity and the type of gating element used (Table III, column 4). For example, gating 50% of the scan flip-flops will yield approximately 42% to 59% reduction in average power consumption in the combinational part, and hence $probX = (1 - 0.5) = 0.5$. For higher values of the required power reduction, the required area overhead tends to deviate more from the required power reduction. For example, 80% area overhead yields an average power reduction of 66% to 77% in the combinational part, including the gating power overhead (Table IV, column 4).

It is worth mentioning that the area overhead value used is the percentage of gated flip-flops with respect to the total number of scan flip-flops, and not the percentage area overhead of gating elements with respect to total circuit area. The latter can be estimated easily, given the gating element size and the total number of scan flip-flops in the circuit.

Without considering the power in the gating elements themselves, the power reduction in the combinational part and the percentage area overhead tend to track closely (Tables III and IV, column 3). This shows that optimizing the gating element can provide better results and that the increased deviation between the two quantities happen primarily as a result of power in the gating elements themselves. The advantage of close correspondence between the two values is that the extra gates area overhead is predictable in a reasonable range, even before computing the details of gating elements placement.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed method, we performed experiments on several ISCAS'89 benchmark circuits (Table I). We used the Synopsys Design Compiler tool [Synopsys 2004a] to perform scan insertion and Synopsys TetraMAX [Synopsys 2004d] for pattern generation. The numbers of scan chains inserted

Table II. Power Reduction Using Full Gating

Benchmark	%ckt avg. w/o GE	%ckt avg. with GE	%ckt peak w/o GE	%ckt peak with GE
s5378	50.72%	38.49%	14.00%	−5.67%
s9234	55.89%	45.44%	7.74%	−4.74%
s13207	43.15%	28.96%	1.87%	−17.89%
s15850	42.63%	28.42%	−8.86%	−30.98%
s35932	42.44%	30.11%	−35.52%	−42.70%
s38417	34.97%	21.45%	6.70%	−17.41%
s38584	38.83%	24.15%	−31.50%	−60.03%

in the benchmark circuits were guided both by the number of flip-flops in the circuit and the desire to keep the size of scan chains in each circuit comparable to others. A single chain was inserted for the first two benchmarks (s5378, s9234), which are relatively small. Three scan chains were inserted in medium-sized benchmarks (s13207, s15850). Eight scan chains were inserted in the larger benchmarks (s35932, s38417, s38584). Note that we did not consider scan ordering in our experiments although we acknowledge that further improvements can be achieved through such ordering.

A stuck-at fault model was used in our experiments. However, scan testing is also applicable for other fault models such as transition delay, path delay, and IDDQ. The ATPG tool was used to write out Verilog testbenches, which in turn were used to simulate the patterns. The choices of gating were evaluated on the scan-inserted netlist using the algorithm of Section 3, and the netlist was edited to include gating elements. We used NOR and OR gates for gating to zero and one (gate-to-0, gate-to-1), respectively. The advantage is that both can use a noninverted scan-enable signal. Adding an inverter after scan-enable can cause a large number of glitches due to difference in arrival times between gate-to-1 and gate-to-0 gates (assuming interconnect delays are balanced). This can greatly increase both average and peak power.

Patterns have been simulated using realistic delay models derived from a physical library. A 180nm technology library was used in our experiments. During pattern simulation, switching activities of the internal nodes were recorded in a separate file. Synopsys PrimePower [Synopsys 2004c] was used to estimate the average and peak test powers. The inputs to PrimePower are the netlist, switching activity file, the target library, which has been characterized for internal power, and pin capacitances. In the simulations, we use a more accurate estimation of power than that used in estimating the cost function, since each cell in the library is characterized for its capacitances, delays, and power dissipation, and a realistic delay model is used. Information related to interconnects was not used in our experiments, since our method is useful in a prephysical design flow.

Table II shows the reduction in average and peak power when all scan flip-flops are gated (full gating method). Column 2 shows the average power reduction when gating overhead is excluded, which makes it independent of the type of gating used. It also shows a high contribution of the power in scan chains in the total circuit power, since the reduction is 38% in the largest circuit even

Table III. Average Power and Peak Power Reduction for 50% Gating Elements

Benchmark	Average Power								Peak Power (with GE)	
	Best Vector (B)			Worst Vector (W)					Best	Worst
	%ckt	%comb	%(comb +GE)	%ckt	%comb	%(comb +GE)	B/W	%(B-W)	%ckt	%ckt
s5378	30.13%	67.52%	59.42%	12.18%	28.48%	24.02%	2.47	35.40%	10.47%	5.87%
s9234	31.15%	61.82%	55.73%	14.61%	29.92%	26.14%	2.13	29.60%	5.69%	7.96%
s13207	23.92%	65.18%	55.43%	10.20%	32.44%	23.65%	2.34	31.78%	−2.94%	−4.56%
s15850	22.52%	62.60%	52.82%	14.95%	43.96%	35.07%	1.51	17.75%	−3.69%	−6.36%
s35932	20.01%	55.79%	47.15%	14.36%	41.35%	33.84%	1.39	13.31%	−20.34%	−11.15%
s38417	15.65%	56.30%	44.75%	10.07%	38.66%	28.78%	1.55	15.97%	−4.96%	−4.09%
s38584	16.49%	53.10%	42.48%	12.74%	42.86%	32.81%	1.29	9.67%	−24.94%	−20.90%

Table IV. Average Power and Peak Power Reduction for 80% Gating Elements

Benchmark	Average Power								Peak Power (with GE)	
	Best Vector (B)			Worst Vector (W)					Best	Worst
	%ckt	%comb	%(comb +GE)	%ckt	%comb	%(comb +GE)	B/W	%(B-W)	%ckt	%ckt
s5378	39.22%	87.81%	77.34%	23.63%	57.23%	46.59%	1.66	30.76%	5.71%	5.23%
s9234	44.83%	89.69%	80.20%	33.26%	67.91%	59.51%	1.35	20.69%	1.56%	2.66%
s13207	32.05%	89.35%	74.28%	19.31%	58.63%	44.74%	1.66	29.54%	−12.11%	−6.87%
s15850	28.26%	80.47%	66.29%	27.46%	77.98%	64.42%	1.03	1.86%	−14.47%	−10.05%
s35932	29.75%	83.12%	70.09%	26.62%	74.91%	62.73%	1.12	7.36%	−33.79%	−25.00%
s38417	23.95%	85.38%	68.49%	17.27%	65.97%	49.37%	1.39	19.12%	−9.98%	−8.37%
s38584	25.92%	83.80%	66.76%	22.09%	72.20%	56.89%	1.17	9.86%	−40.35%	−32.95%

when the combinational part is totally turned off. This can be a direct result of the particular library we used (such as the high clock pin input capacitance). Layout optimizations and buffering can help to reduce this power, independent of the use or nonuse of gating, and the type of gating elements. Column 3 shows the average power reduction when power in the gating elements is considered. It shows that significant power is consumed in the gating elements themselves, especially when the number of gating elements is large, as in full gating. In all of the benchmarks, the peak power increased from about 5% and up to 60% when the gating elements overhead is considered. As pointed out earlier, this increase is due to a large number of gating elements changing state, either when the scan chains change from shift mode to capture mode or during capture itself.

Tables III and IV show the results when the number of gating elements is 50% and 80% of the total number of scan flip-flops, respectively. In each table, we show the reduction in average power and peak power when: (a) best vector and (b) worst vector was used. The best and worst vectors were found using the INPUT_ASSIGNMENT procedure explained in the previous section. We experimented with different numbers of the total number of iterations from 10,000 to 1,000,000. A slight improvement has been observed going from 10,000 to 1,000,000. This involves a tradeoff between running time and result improvement. It has also been observed that sometimes a slight improvement in the cost function may translate into a slight increase in power. This is due to the inherent inaccuracy of the fanout weighting function versus the more accurate

estimation methods of the power estimation tool. Entries under the %ckt column are the power reductions in the entire circuit, whereas the entries under %comb are power reductions in the combinational part only, and %(comb+GE) is the power reduction in both combinational logic and gating elements. In Table II (full gating), we have not included a column for %comb since there is no switching activity in the combinational part during scan-in. Specifically, %comb serves to give a measure that is independent of both the power in scan chains, and the type of gating element used. As discussed before, both can be subject to further optimizations.

Based on Tables III and IV, we make the following observations:

—For 50% gating, the ratio of the average power reduction in the best and worst vectors ranges from 1.29 to 2.47. The difference tends to narrow as gating percentage increases.

—When 50% of the flip-flops are gated, the achieved reduction in average power is more than 50% of the reduction when all flip-flops are gated, even when gating overhead is included.

—If we compare column 3 in Table II and column 2 in Table IV, we can see that when gating overhead is included, the total reduction in average power is almost the same as the reduction when full gating is used. This shows that with less area overhead, we can achieve almost the same average power reduction and have less impact on peak power. Comparison of columns 5 in Table II and 10 in Table IV shows that while peak power increases in both cases, the effect of full gating is worse. This slight difference occurs because the extra gating elements in full gating consume more power when switching from blocking to transparent mode, or after capture.

—Comparing Tables III and IV, we find that the average power reduction in the combinational part always exceeds the gating percentage if gating points are properly selected, but when gating overhead is included, the gating overhead starts to offset the savings achieved by extra gating. As the gating percentage is increased, the power in the gating elements themselves goes higher.

—Gating overhead can have considerable impact on both average and peak power, which suggests the use of special types of gates as gating elements or the use of low-overhead gating techniques, such as the one in Bhunia et al. [2004], that will reduce their impact. In practice, the regular NOR/OR gate that we used in our simulations is not the ideal choice from a power, area, or performance perspective.

These observations clearly indicate that careful selection of gating elements can prove to be very effective in test power reduction, and optimization of the gating elements themselves is necessary to reduce their impact on power savings.

The area increase due to 50% partial and full gating is shown in Table V. The area advantage of partial gating is clear from the table. Although not shown in the table, the area savings in interconnect and routing can be significant.

Since added gates will have an effect on the power consumption during normal mode operation, some general considerations have been applied in regard

Table V. Gating Area Overhead

Benchmark	50% gating	Full gating
s5378	2.25%	3.41%
s9234	1.76%	2.76%
s13207	2.47%	4.34%
s15850	2.09%	3.58%
s35932	2.97%	5.09%
s38417	2.43%	4.20%
s38584	2.19%	4.04%

Table VI. Normal Mode Power Consumption Increase with Gating

Benchmark	Average Power		Peak Power	
	50% Gating	Full Gating	50% Gating	Full Gating
s5378	4.85%	7.94%	21.51%	28.49%
s9234	2.77%	6.83%	4.78%	14.84%
s13207	2.64%	6.76%	3.54%	15.02%
s15850	3.69%	7.06%	7.09%	20.69%
s35932	2.47%	4.24%	2.41%	4.72%
s38417	1.00%	2.46%	2.03%	4.99%
s38584	0.04%	0.10%	0.00%	0.00%

to the way gating elements operate when they are in their transparent mode. For a NOR gate, less capacitance is charged/discharged during switching if the input connected to the test-enable signal is farther from the output. Since this input will be only fixed at zero during normal mode, the capacitance of the associated transistor does not need to charge/discharge every time the other input switches. Also if the gating elements are not placed on the critical paths of the circuit and extra delay of the gating elements can be tolerated, the gating elements can be sized smaller so that they have less capacitance. Also gates of higher threshold can be used on noncritical paths. Table VI shows a comparison between the power dissipated in normal mode for full gating, 50% partial gating, and no gating for the benchmark circuit s5378. Since functional patterns are not available for the benchamark circuits, we generated 1,000,000 random patterns with 25% of the transitions between consecutive patterns applied to primary inputs.

5. PEAK POWER HANDLING

As pointed out earlier, a drawback of the gating technique is that while it reduces average test power, it may increase the peak test power. Excessive switching happens when the gating elements change from their fixed status to their transparent mode. In this event, switching propagates to the combinational logic gates, causing many gates to switch at once and the subset of gating elements that will switch changing all at the same time, especially when the number of gating elements is large (as in full gating). The same happens during capture when all flip-flops change from present state to next captured state along with all the following logic. Again, the existence of gating elements causes the peak power to increase when they switch almost at the same time after capture, with extra glitching activities caused by imbalances in their paths from the scan cell outputs.

Table VII. Peak Power Reduction with SeqTD and BC for 50% Gating Elements

Benchmark	Partial Gating only	Partial Gating and SeqTD	Partial Gating, SeqTD, and BC
s13207	−2.94%	−2.94%	13.10%
s15850	−3.69%	−3.99%	8.29%
s35932	−20.34%	−6.72%	5.52%
s38417	−4.96%	−4.96%	8.10%
s38584	−24.94%	0.76%	12.43%

Instead of switching all the gating elements at once, we can do this process in stages. Since medium to large circuits usually have more than one scan chain, we can control scan chains one-at-a-time for test enabling/disabling and capture. In such a case only a partial number of gates switches at once, and peak power is not excessive. Since the number of scan chains is much smaller than the number of flip-flops in a scan chain, the impact on test time is minimal.

5.1 Sequential Test Disable (SeqTD)

If the peaks occur primarily as a result of gating elements switching at once from blocking to transparent modes, multiple test-enable signals can be provided that activate and deactivate sequentially. Now, only a subset of scan cells and a subset of gating elements will switch at the same time and the peak currents at this instance will go lower. The effect on average power is minimal. This technique is different from those proposed in Sankaralingam et al. [2001], Whetsel [2000], Nicolici and Al-Hashimi [2000], Saxena et al. [2001], and Ghosh et al. [2003] in that all scan chains will do the shifting together, while the sequencing happens only with the enable and disable of the test-enable signals, saving test time considerably.

Table VII shows comparison of peak power reduction for the last five benchmark circuits when 50% gating is used (column 2) and when SeqTD is used (column 3). We can see that in two out of the five benchmark circuits, the peak power has improved considerably, while in three of them, either no or slight change has been observed. Those circuits with no change in peak power have their peak power caused mainly by capture. An added advantage is that not only the peak power is improved, but also the instances at which peak currents occur. Figures 6 and 7 show how the power profile improved considerably by using SeqTD.

5.2 Blocked Capture (BC)

If peak power occurs during capture cycles, then we attempt to block captured signals from propagating to combinational logic and gating elements, since neither of them are needed in capture mode. We can achieve this by placing either a latch or a tri-state buffer before each gating element, as shown in Figure 8. Before the capture clock is applied, the capture-blocking gate is disabled and remains so until after scan-in of the next pattern is completed and when signals are allowed to propagate through combinational logic. This guarantees that captured signal effects will only be localized to the scan flip-flops, while the gating elements remain functioning normally so as to inhibit transitions
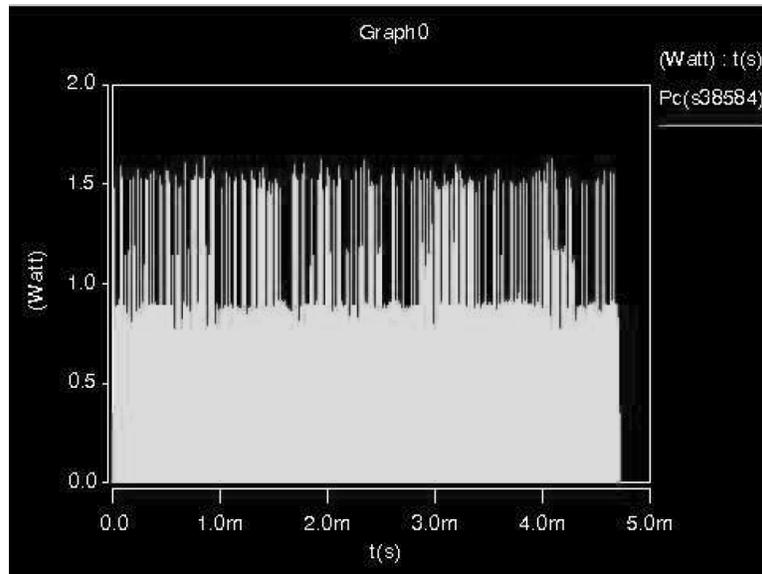
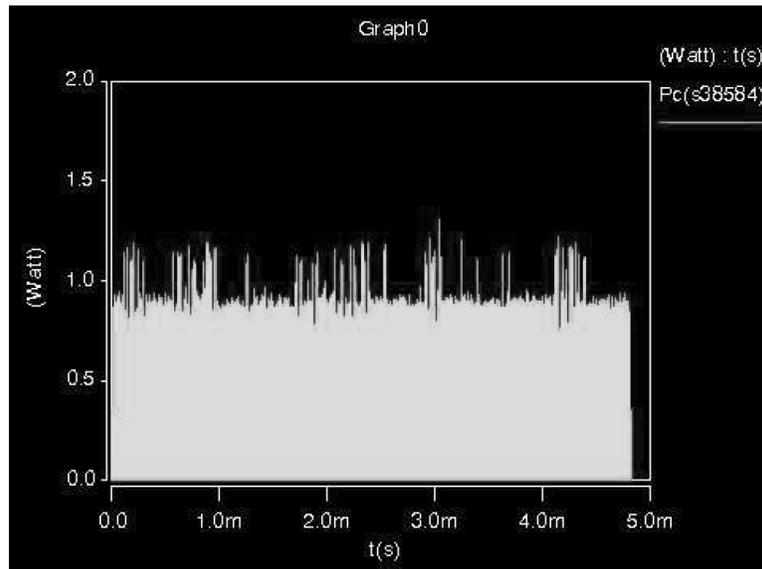Fig. 6.   Power profile before SeqTD.

Fig. 7.   Power profile after SeqTD.

in combinational logic during scan-in. The sequence of operations is shown in Figure 9 in which blocked capture is combined with sequential test disable (SeqTD). Captured signals are only allowed to propgate once test enable is reenabled. Since this is done sequentially, there is no peak power during capture. It is clear that no data dependency problems occur since all scan cells are captured at the same time.
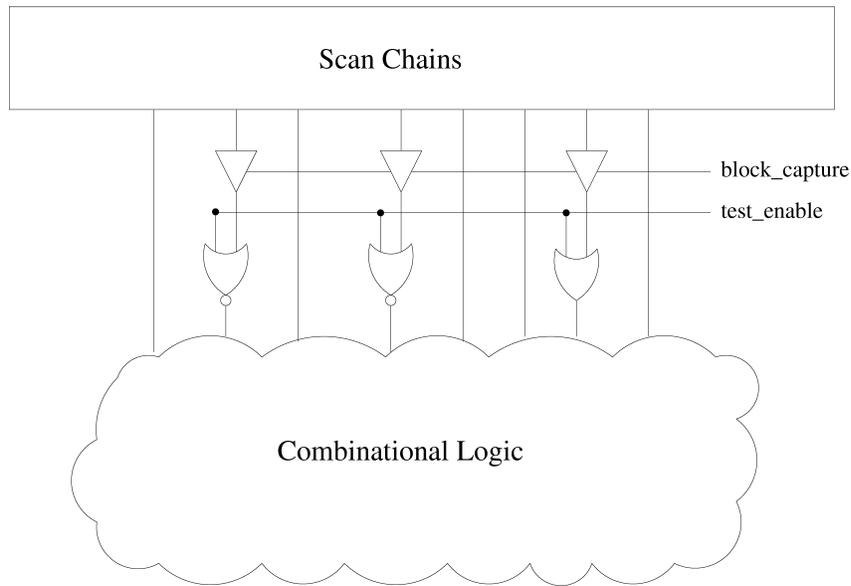
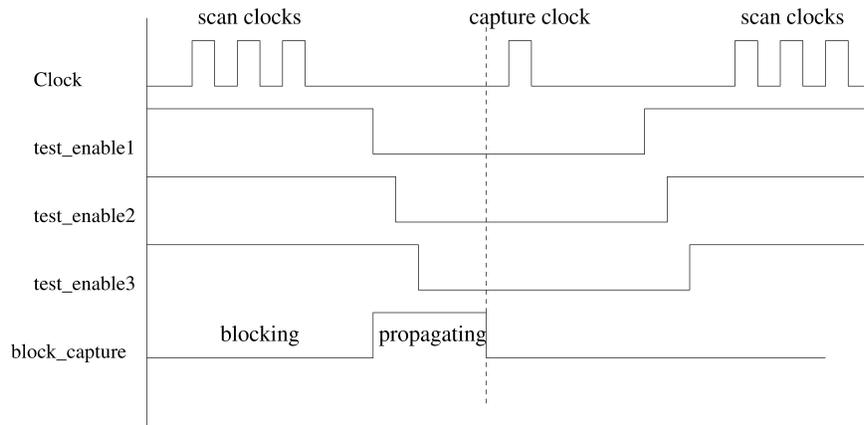Fig. 8.   Partial gating with blocked capture (BC).

Fig. 9.   Waveform sequence for SeqTD and BC.

To demonstrate the advantage of blocked capture combined with sequential test disable, even in the worst-case scenario where a latch and a NOR/OR gate is placed after each gated scan cell, we report the results in the last column of Table VII. It is clear that even with the overhead that these elements introduced, the savings in peak power is still noticeable.

5.3 Delay Fault Testability

Adding the aforementioned techniques for reducing the impact of peak power can raise some concerns regarding delay fault testability. In this subsection we

touch upon this subject. First, if SeqTD is used alone, delay fault testing can be done using the launch-off-capture (LOC) technique. In such a case, two clocks are applied after the initial test pattern scan-in. Sequencing of the test-enable signals will not affect the delay testing function because it only applies before the first capture clock, whereas test-enable signals remain disabled until after the second at-speed capture clock. Launch-off-shift (LOS) techniques cannot be used with this method because of the conflicting requirement that enable signals are to be turned off immediately after the shift.

If BC is used along with SeqTD, then the only way that delay fault testing can be done is by using enhanced scan [Bushnell and Agrawal 2000]. In that case the blocking gates will be used to hold the vector to be applied until the initializing vector is scanned-in. Clearly the blocking gates in that case should be applied to all scan flip-flops and not partially.

6. CONCLUSION

Other than scan data volume reduction, which has received a lot of attention in the recent past, another issue that has generated a great deal of interest is scan test power reduction. Full gating techniques, which authors have proposed in the past so as to convert all the scan flip-flops to gated scan flip-flops, is area-intensive, complex from a physical design viewpoint, and intrusive from the timing closure viewpoint. The main contribution of this article is to take advantage of structural properties of the circuit to avoid full gating. A user-specified percentage of scan cells can be converted to gated scan cells in the proposed methodology, allowing the user to control the overheads of scan gating. An algorithm was used to select scan cells for gating. In our results, even 50% gating gave up to 30% average power reduction where a 100% gating provides about 38% reduction. We showed that the overhead in gating elements themselves can have serious effects, especially on peak power, and cannot be ignored. We also showed techniques to handle the excessive increase in peak power caused by gating. We provided results that are independent of both the power consumed in the scan chain and the type of gating elements used.

REFERENCES

AHMED, N., TEHRANIPOOR, M., AND NOURANI, M. 2004. Low power pattern generation for BIST architecture. Vol. II. In *Proceedings of the International Symposium on Circuits and Systems (ISCAS)*. 689–92.

BHUNIA, S., MAHMOODI, H., MUKHOPADHYAY, S., GHOSH, D., AND ROY, K. 2004. A novel low-power scan design technique using supply gating. In *Proceedings of the International Conference on Computer Design: VLSI in Computers and Processors (ICCD)*. 60–65.

BUSHNELL, M. L. AND AGRAWAL, V. D. 2000. *Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits*. Kluwer Academic, Boston, MA.

BUTLER, K., SAXENA, J., JAIN, A., FRYARS, T., LEWIS, J., AND HETHERINGTON, G. 2004. Minimizing power consumption in scan testing: Pattern generation and DFT techniques. In *Proceedings of the International Test Conference (ITC)*. 355–364.

CHAKRAVARTY, S. AND DABHOLKAR, V. 1994. Two techniques for minimizing power dissipation in scan circuits during test application. In *Proceedings of the Asian Test Symposium (ATS)*. 324–329.

CHOU, R., SALUJA, K., AND AGRAWAL, V. 1997. Scheduling tests for VLSI systems under power constraints. *IEEE Trans. Very Large Scale Integration Syst. 5,* 2, 175–185.

DABHOLKAR, V., CHAKRAVARTY, S., POMERANZ, I., AND REDDY, S. 1998. Techniques for minimizing power dissipation in scan and combinational circuits during test application. *IEEE Trans. Comput. Aided Des. Integrated Circ. Syst. 17,* 12, 1325–1333.

GERSTENDORFER, S. AND WUNDERLICH, H.-J. 1999. Minimized power consumption for scan-based BIST. In *Proceedings of the International Test Conference (ITC)*. 77–84.

GHOSH, D., BHUNIA, S., AND ROY, K. 2003. Multiple scan chain design technique for power reduction during test application in BIST. In *Proceedings of the IEEE International Symposium on Defect and Fault Tolerance in Very Large Scale Integrated Systems*. 191–198.

GIRARD, P. 2002. Survey of low-power testing of VLSI circuits. *IEEE Des. Test Comput. 19,* 3, 80–90.

GIRARD, P., GUILLER, L., LANDRAULT, C., AND PRAVOSSOUDOVITCH, S. 1999a. Circuit partitioning for low power BIST design. In *Proceedings of the Asian Test Symposium (ATS)*. 89–94.

GIRARD, P., GUILLER, L., LANDRAULT, C., AND PRAVOSSOUDOVITCH, S. 1999b. A test vector ordering technique for switching activity reduction during test operation. In *Proceedings of the Great Lakes Symposium on Very Large Scale Integration*. 24–27.

GIRARD, P., GUILLER, L., LANDRAULT, C., AND PRAVOSSOUDOVITCH, S. 2000. Low power BIST design by hypergraph partitioning: Methodology and architectures. In *Proceedings of the International Test Conference (ITC)*. 652–661.

GIRARD, P., LANDRAULT, C., PRAVOSSOUDOVITCH, S., AND SEVERAC, D. 1998. Reducing power consumption during test application by test vector ordering. In *Proceedings of the International Symposium on Circuits and Systems (ISCAS)*. 296–299.

HUANG, T. AND LEE, K. 2001. Reduction of power consumption in scan-based circuits during test application by an input control technique. *IEEE Trans. Comput. Aided Des. Integrated Circ. Syst. 20,* 7, 911–917.

NICOLICI, N. AND AL-HASHIMI, M. 2000. Scan latch partitioning into multiple scan chains for power minimization in full scan sequential circuits. In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE)*. 715–722.

RAVIKUMAR, C., CHANDRA, G., AND VERMA, A. 2000. Simultaneous module selection and scheduling for power-constrained testing of core based systems. In *Proceedings of the International Conference on Very Large Scale Integration Design*. 462–467.

SANKARALINGAM, R., ORUGANTI, R., AND TOUBA, N. 2000. Static compaction techniques to control scan vector power dissipation. In *Proceedings of the Very Large Scale Integrated Test Symposium (VTS)*. 35–40.

SANKARALINGAM, R., POUYA, B., AND TOUBA, N. 2001. Reducing power dissipation during test using scan chain disable. In *Proceedings of the Very Large Scale Integrated Test Symposium (VTS)*. 319–324.

SANKARALINGAM, R. AND TOUBA, N. 2002a. Controlling peak power during scan testing. In *Proceedings of the Very Large Scale Integrated Test Symposium (VTS)*. 153–159.

SANKARALINGAM, R. AND TOUBA, N. 2002b. Inserting test points to control peak power during scan testing. In *Proceedings of the International Symposium on Defect and Fault Tolerance in Very Large Scale Integrated Systems (DFT)*. 138–146.

SAXENA, J., BUTLER, K., AND WHETSEL, L. 2001. An analysis of power reduction techniques in scan testing. In *Proceedings of the International Test Conference (ITC)*. 670–677.

SYNOPSYS INC. 2004a. *Synopsys Design Compiler User Manual for Synopsys Toolset version 2004.06*. Synopsys Inc.

SYNOPSYS INC. 2004b. *Synopsys DFT Compiler User Manual for Synopsys Toolset version 2004.06*. Synopsys Inc.

SYNOPSYS INC. 2004c. *Synopsys Prime Power User Manual for Synopsys Toolset version 2004.06*. Synopsys Inc.

SYNOPSYS INC. 2004d. *Synopsys TetraMax User Manual for Synopsys Toolset version 2004.06*. Synopsys Inc.

WANG, S. AND GUPTA, S. 1994. ATPG for heat dissipation minimization during test application. In *Proceedings of the International Test Conference (ITC)*. 250–258.

WANG, S. AND GUPTA, S. 1997. ATPG for heat dissipation minimization during scan testing. In *Proceedings of the Design Automation Conference (DAC)*. 614–619.

WHETSEL, L. 2000. Adapting scan architectures for low power operation. In *Proceedings of the International Test Conference (ITC)*. 863–872.

ZHANG, X. AND ROY, K. 2000. Power reduction in test-per-scan BIST. In *Proceedings of the IEEE On-Line Testing Workshop*. 133–138.

ZORIAN, Y. 1993. A distributed BIST control scheme for complex VLSI devices. In *Proceedings of the Very Large Scale Integrated Test Symposium (VTS)*. 4–9.