Combining Low-Power Scan Testing and Test Data Compression for System-on-a-Chip*

Anshuman Chandra and Krishnendu Chakrabarty
Dept. Electrical and Computer Engineering
Duke University
Durham, NC 27705, USA.
{achandra, krish}@ee.duke.edu

ABSTRACT

We present a novel technique to reduce both test data volume and scan power dissipation using test data compression for system-on-a-chip testing. Power dissipation during test mode using ATPG-compacted test patterns is much higher than during functional mode. We show that Golomb coding of precomputed test sets leads to significant savings in peak and average power, without requiring either a slower scan clock or blocking logic in the scan cells. We also improve upon prior work on Golomb coding by showing that a separate cyclical scan register is not necessary for pattern decompression. Experimental results for the larger ISCAS 89 benchmarks show that reduced test data volume and low power scan testing can indeed be achieved in all cases.

Keywords

Embedded core testing, Golomb codes, precomputed test sets, scan testing, switching activity, test set encoding.

1. INTRODUCTION

Pre-designed intellectual property (IP) cores are now commonly used in large system-on-a-chip (SOC) designs. However, IP cores pose several difficult test challenges. Two problems that are becoming increasingly important are power consumption during manufacturing test and test data volume. The precomputed test patterns provided by the core vendor must be applied to each core within the power constraints of the SOC. In addition, test data compression is necessary to overcome the limitations of the automatic test equipment (ATE), e.g. tester data memory and I/O channel capacity.

Power consumption during testing is important since excessive heat dissipation can damage the circuit under test. Since power consumption in test mode is higher than during normal operation, special care must be taken to ensure that the power rating of the SOC is not exceeded during test application [1]. A number of techniques to control power consumption in test mode have been presented in the literature. These include test scheduling algorithms under power constraints [2], low-power built-in self-test (BIST) [3, 4], and techniques for minimizing power during scan testing [5, 6, 7]. Power consumption is especially important for SOCs since test scheduling techniques for system integration attempt to reduce testing time by applying scan/BIST vectors to several cores simultaneously [8, 9]. Therefore, it is extremely important to decrease power consumption while testing the IP cores in an SOC.

Test data volume is another problem faced in SOC test integration. One way to alleviate this problem is to use BIST. However, BIST can only be applied to SOCs if the IP cores in them are BIST-ready. Since most currently-available IP cores are not BIST-ready, the incorporation of BIST in them requires considerable redesign. Hence test data compression techniques that facilitate low-power scan testing are desirable for SOC testing.

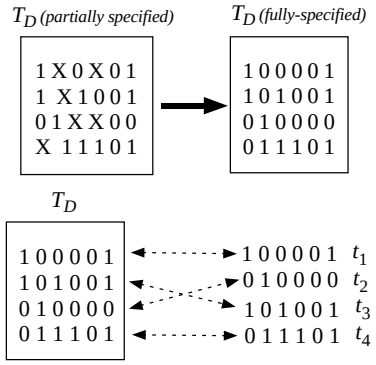
The conflicting goals of low-power scan testing and reduced test data volume appear to be irreconcilable. Test generation for low-power scan testing usually leads to an increase in the number of test vectors [5]. On the other hand, static compaction of scan vectors causes significant increase in power consumption during testing [7]. The compacted vectors are rendered useless if they exceed power constraints. Clearly, uncompacted vectors cannot be used since they require excessive tester memory. This problem is addressed in a recent paper on power-constrained static compaction of scan vectors [7]. However, while [7] provides 2-3 times reduction in power consumption for several ISCAS benchmark circuits, it does not lead to any appreciable reduction in test data volume—in fact, it does not provide any improvement over standard static vector compaction techniques.

Recently, a number of data compression techniques have been proposed for reducing SOC test data volume [10, 11]. In this approach, the precomputed test set T_D provided by the core vendor is compressed (encoded) to a much smaller test set T_E and stored in ATE memory. An on-chip decoder is used for pattern decompression to generate T_D from T_E during pattern application. It was shown in [10, 11] that compressing a "difference vector" sequence T_{diff} determined from T_D results in smaller test sets and reduces testing time. An obvious drawback of this approach is that it requires a separate cyclical scan register (CSR).

In this paper, we dispel the notion that scan vector compaction always leads to higher power consumption. Since static compaction invariably leads to higher power, we ex-

*This research was supported in part by the National Science Foundation under grant number CCR-9875324.

DAC 2001 June 18-22, 2001, Las Vegas, Nevada, USA.



$T_D = \{t_1, t_2, t_3, t_4\} = \{100001, 010000, 101001, 011101\}$

Figure 1: An example to illustrate the procedure of deriving fully-specified ordered T_D.

plore test data compression for overcoming this problem. We show that we can decrease both peak and average power by using Golomb codes for compressing the scan vectors of IP cores. In addition, we show that it is not necessary to use a separate CSR; we can directly encode T_D.

2. COMPRESSION METHOD AND TEST ARCHITECTURE

We first review Golomb coding and its application to test data compression in [11]. If the difference vector T_{diff} is used for compression, the first step is to derive it from T_D, where $T_D = \{t_1, t_2, t_3, ..., t_n\}$, is the (ordered) precomputed test set. The ordering is determined using a heuristic procedure described [11]. T_{diff} is defined as follows:

$$T_{diff} = \{d_1, d_2, \ldots, d_n\} = \{t_1, t_1 \oplus t_2, t_2 \oplus t_3, ..., t_{n-1} \oplus t_n\}.$$

where a bit-wise exclusive-or operation is carried out between patterns t_i and t_j.

In this work however, we encode T_D directly, hence there is no need to generate T_{diff}. All the don't-care bits in T_D are mapped to 0s to obtain a fully-specified test sequence.

The next step in the encoding procedure is to select the Golomb code parameter m, referred to as the group size. Once m is determined, the runs of 0s in the test data stream are mapped to groups of size m (each group corresponding to a run length). The mapping procedure for obtaining the codewords is described in [11].

The problem of determining the best ordering is equivalent to the NP-Complete Traveling Salesman problem. Therefore, a greedy algorithm is used to generate an ordering and the corresponding T_D. Suppose a partial ordering $t_1 t_2 \ldots t_i$ has already been determined for the patterns in T_D. To determine t_{i+1}, we calculate the Hamming distance $HD(t_i, t_j)$ between t_i and all patterns t_j that have not been placed in the ordered list. We define $HD(t_i, t_j)$ as the number of 0s in the pattern t_j. We select the pattern t_j for which $HD(t_i, t_j)$ is maximum and add it to the ordered list, denoting it by t_{i+1}. In this way, a fully-specified test pattern is obtained and the smallest number of 1s is added to the ordered vector sequence. We continue this process until all test patterns in T_D are placed in the ordered list. Figure 1 illustrates the procedure for obtaining fully specified ordered T_D.

An on-chip decoder decompresses the encoded test set T_E and produces T_D. Even though T_D contains more patterns than test sets obtained after static compaction of ATPG vectors, the testing time is reduced since pattern decompression can be carried out on-chip at higher clock frequencies. As discussed in [11], the decoder can be efficiently implemented by a $\log_2 m$-bit counter and a finite-state machine (FSM). The synthesized decode FSM circuit contains only 4 flip-flops and 34 combinational gates. For any circuit whose test set is compressed using $m = 4$, the given logic is the only additional hardware required other than the 2-bit counter. This is especially the case if, unlike in [11], T_D is directly used for encoding and a CSR is not required for decompression.

Since the decoder for Golomb coding needs to communicate with the tester, and both the codewords and the decompressed data can be of variable length, proper synchronization must be ensured through careful design. In particular, the decoder must communicate with the tester to signal the end of a block of variable-length decompressed data. These and other related decompression issues are discussed in detail in [11].

3. POWER ESTIMATION FOR SCAN VECTORS

In this section, we examine the impact of test set encoding on power consumption during scan testing. We then show how power consumption can be minimized by appropriately assigning binary values to the don't-care bits in T_D and then applying Golomb coding for test data compression.

For a CMOS circuit, power consumption can be classified as either static or dynamic. Static power consumption, which is caused by leakage current, is usually negligible and therefore ignored. Dynamic power is consumed when the the outputs of circuit elements from high-to-low and low-to-high transitions. This constitutes the predominant fraction of CMOS power consumption.

For scan vectors, the dynamic power consumption during testing depends on the number of transitions that occur in the scan chain as well as on the number of circuit elements that switch during the scan in and scan out operations. It is difficult to estimate the scan out power directly from the scan vector set since the test responses must be determined from the function of the core under test. Therefore, as in [7], we limit ourselves to the scan in power only and measure it in terms of the number of transitions in the scan vectors. We also use the weighted transitions metric introduced in [7] to estimate the power consumption due to scan vectors. This models the fact that the scan in power for a given vector depends not only on the number of transitions in it but also on their relative positions. For example, consider a scan vector $v_1 v_2 v_3 v_4 v_5 = 01000$, where v_1 is first loaded into the scan chain. The 0-to-1 transition between v_1 and v_2 causes more switching activity in the scan chain than the 1-to-0 transition between v_2 and v_3.

The weighted transitions count metric is also strongly correlated to the switching activity in the internal nodes of the core under test during the scan in operation. It was shown experimentally in [7] that scan vectors have have higher weighted transition metric dissipate more power in the core under test.

Consider a scan chain of length l and a scan vector $t_j = t^\star_{j,1} t^\star_{j,2} \ldots t^\star_{j,l}$, with $t^\star_{j,1}$ scanned in before $t^\star_{j,2}$, and so on. The *weighted transitions metric* for t_j, denoted WTM_j, is given by $WTM_j = \sum_{i=1}^{l-1} (l-i) \cdot (t^\star_{j,i} \oplus t^\star_{j,i+1})$. If the test set

Partially-specified scan vector	Fully-specified vector (Minimum WTM)	Fully-specified vector (Don't-cares mapped to 0s)
$t_i = 01XXX10XXX01$	011111000001 Golomb code length: 19 bits ($m = 4$) $WTM_i = 18$	010001000001 Golomb code length: 10 bits ($m = 4$) $WTM_i = 37$
$t_j = 0XX1010XXXX1$	000101000001 Golomb code length: 10 bits ($m = 4$) $WTM_j = 31$	010101000001 Golomb code length: 13 bits ($m = 4$) $WTM_j = 52$

Table 1: Mapping of don't-cares in T_D to binary values.

T_D contains n vectors $t_1, t_2, \ldots, t_n$ then the average scan in power P_{avg} and peak scan in power P_{peak} are estimated as follows:

$$P_{avg} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{l-1}(l-i)\cdot(t^{\star}_{j,i} \oplus t^{\star}_{j,i+1})}{n}$$

$$P_{peak} = \max_{j\in\{1,2,\ldots,n\}}\{\sum_{i=1}^{l-1}(l-i)\cdot(t^{\star}_{j,i} \oplus t^{\star}_{j,i+1})\}.$$

If the peak power exceeds a threshold value, it can cause structural damage to the silicon or to the package. Likewise, elevated average power can also cause structural damage to the silicon, bonding wires or the package. It also adds to the thermal load that must be transported away from the device under test.

We next show how Golomb codes can be used to minimize the volume of test data and at the same time, minimize P_{avg} and P_{peak}. Scan-in power is influenced by the manner in which the don't-cares in T_D are mapped to binary values. While P_{avg} and P_{peak} can be minimized by choosing an appropriate mapping, such a mapping is not guaranteed to provide high test data compression. In fact, our experiments show that the encoded test sets in such cases are often larger than the uncompacted test sets. Instead, it is far more efficient to simply map all the don't-cares in T_D to 0s as shown in Figure 1. While this approach does not minimize P_{avg} and P_{peak}, it provides significant reductions in power consumption, and at the same time, decreases the test data volume considerably. The fully-specified test set thus obtained is then compressed using Golomb codes.

For example, Table 1 shows two partially-specified scan vectors $t_i = 01---1----01$ and $t_j = 01-1010----1$ with scan chain length $l = 12$, where $-$ denotes a don't-care bit. If the don't-cares are mapped to binary values to minimize the weighted transition metric, then $d---d'$, $d \in \{0,1\}$, must be mapped to $dddddd'$. Similarly, $d----$ must be mapped to $ddddd$. This ensures that the few unavoidable transitions occur "late" during scan in. Table 1 shows the values of WTM_i and WTM_j and the Golomb codes for the corresponding fully-specified vectors ($m = 4$). The weighted transitions metric is clearly higher if the don't-cares are always mapped to 0. However, Golomb coding is much more effective in reducing test data volume if this strategy is used.

Next we present the following theorem which characterizes the maximum WTC for a given test length n, scan chain length l, and the number of 1s (r) in the test set. The proof is omitted for conciseness. This yields the maximum value for the average power P_{avg} and it can be used to predict average power by using limited information about T_D. The maximum value for the peak power is easier to derive—it simply equals $l(l+1)/2$ as long as $r \geq l/2$.

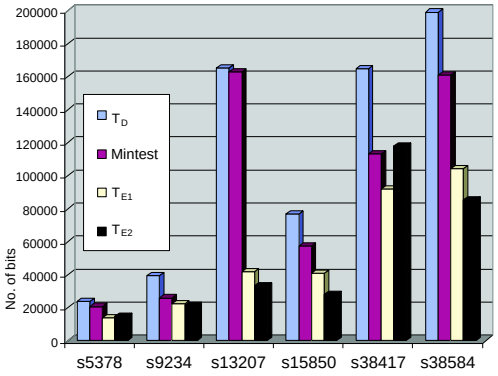


Figure 2: Experimental results on test data compression using Golomb codes.

THEOREM 1. *For a given test length n, scan chain length l, and the number of 1s r in the test set, an upper bound on the average power is given by*

$$P_{avg} \leq \frac{lr}{n} - \frac{r^2}{n^2} + \frac{r}{2n^3}(\frac{r}{n}+1).$$

4. EXPERIMENTAL RESULTS

In this section, we evaluate the effect of Golomb coding of T_D on test data volume and power consumption during scan testing for the ISCAS 89 benchmark circuits. The experiments were conducted on a Sun Ultra 10 workstation with a 333 MHz processor and 256 MB of memory. We only considered the large full-scan circuits with a single scan chain each. The test vectors for these circuits were reordered to increase compression.

Figure 2 presents the experimental results for test cubes T_D obtained from the Mintest ATPG program with dynamic compaction. In order to compare with [11], we also present compressed results obtained using the difference vector sequences T_{diff} (T_{E1}) for the same test sets. Figure 2 shows the sizes of T_D, the size of the smallest encoded test set obtained after static compaction using Mintest, size of T_{E1} and the size of compressed test set obtained using T_D with all Xs mapped to 0 (T_{E2}).

As is evident from Figure 2, T_D yields better compression than T_{diff} in four out of the six cases. For these circuits, we achieve better compression without requiring a separate CSR. Therefore, there is a significant reduction in hardware overhead as compared to [11]. The results also show that ATPG compaction may not always be necessary for saving memory and reducing testing time. In five out of the six cases, the size of the encoded test set is less than the smallest ATPG-compacted test sets known for these circuits. This comparison is essential in order to show that storing T_E in ATE memory is more efficient than simply applying static compaction to test cubes and storing the resulting compact test sets. On average, the size of T_E is 36.26% less than that of the compacted test sets obtained using Mintest.

We next present results on the peak and average power consumption during the scan-in operation. These results show that test data compression can also lead to significant savings in power consumption. As described in Section 3, we estimate power using the weighted transitions metric. Let

Circuit	Uncompacted test sets with don't-cares mapped to 0s				Uncompacted test sets with don't-cares mapped to minimize WTM			
	Peak power P^G_{peak}	Peak power reduction (percent)	Average power P^G_{avg}	Average power reduction (percent)	Peak power P^G_{peak}	Peak power reduction (percent)	Average power P^G_{avg}	Average power reduction (percent)
s5378	10127	24.55	3336	69.89	9531	28.99	2435	78.02
s9234	12994	25.72	5692	61.09	12060	31.06	3466	76.30
s13207	101127	25.42	12416	89.82	97606	28.02	7703	93.68
s15850	81832	18.35	20742	77.18	63478	36.66	13381	85.27
s35932	172834	75.56	73080	87.47	125490	82.25	46032	92.11
s38417	505295	26.10	172665	71.31	404617	40.82	112198	81.35
s38584	531321	7.21	136634	74.50	479530	16.25	88298	83.52
Average	—	28.98	—	75.89	—	37.72	—	84.32

Table 2: Impact of the mapping of don't-cares to binary values on power consumption.

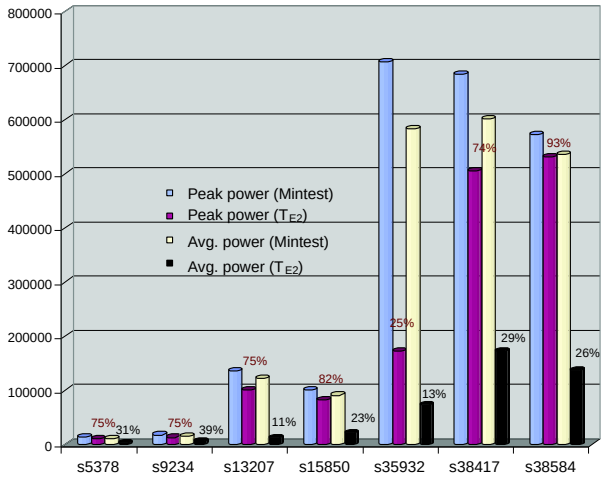


Figure 3: Experimental results on peak and average power consumption.

P^C_{peak} (P^C_{avg}) be the peak (average) power with compacted test sets obtained using Mintest. Similarly, let P^G_{peak} (P^G_{avg}) be the peak (average) power when Golomb coding is used by mapping the don't-cares in T_D to 0s. Figure 3 compares the average and peak power consumption for Mintest test sets with T_D when Golomb coding is used. The percentage reduction in power was computed as follows:

Figure 3 shows that the peak power and average power are significantly less if Golomb coding is used for test data compression and the decompressed patterns are applied during testing. On average, the peak (average) power is 28.98% (75.89%) less in this case than for the Mintest test sets. Thus our results demonstrate that the substantial reduction in test data volume is also accompanied by significant reduction in power consumption during scan testing.

Next, we justify the the strategy of mapping all don't-cares in T_D to 0s before Golomb coding. As discussed in Section 3, the power consumption can be minimized if the don't-cares are assigned to binary values to minimize the weighted transitions metric. Unfortunately, this strategy does not lead to any significant decrease in the test data volume—in fact, we found that in many cases, the encoded test set was larger than the original test set. We therefore carried out a set of experiments to demonstrate that if all don't-cares are mapped to 0s, the test data volume decreases substantially (Table 2) and at the same time, power savings are significant.

Our experimental results for the larger ISCAS 89 circuits are shown in Table 2. We note that while the average power consumption is greater compared to the "optimal" mapping of don't-cares, it is still significantly less than the power for ATPG-compacted test sets. In some cases, the difference is as low as 4%, while on average, the average power consumption increases by only 8%. Likewise, the difference in peak power consumption is only 9% on average. Nevertheless, compared to Mintest, we achieve 51% test data compression on average with 76% reduction in average power consumption for scan testing. This provides a strong justification for the proposed test data compression approach.

5. REFERENCES

[1] Y. Zorian, "A distributed BIST control scheme for complex VLSI devices", *Proc. VTS*, pp. 4-9, 1993.

[2] R. M. Chou, K. K. Saluja and V. D. Agarwal, "Scheduling tests for VLSI systems under power constraints", *IEEE Trans. on VLSI Systems*, vol. 5, pp. 175–185, June 1997.

[3] S. Gerstendörfer and H.-J. Wunderlich, "Minimized power consumption for scan-based BIST", *Proc. ITC*, pp.77-84, 1999.

[4] P. Girard, L. Guiller, C. Landrault and S. Pravossoudovitch, "A test vector inhibiting technique for low energy BIST design", *Proc. VTS*, pp. 407-412, 1999.

[5] S. Wang and S. K. Gupta, "ATPG for heat dissipation minimization during scan testing", *Proc. DAC*, pp. 614-619, 1997.

[6] V. Dabholkar, S. Chakravarty, I. Pomeranz and S. M. Reddy, "Techniques for minimizing power dissipation in scan and combinational circuits during test application", *IEEE Trans. on CAD*, vol. 17, No. 12, pp. 1325-1333, Dec. 1998.

[7] R. Sankaralingam, R. R. Oruganti and N. A. Touba, "Static compaction techniques to control scan vector power dissipation", *Proc. VTS*, pp. 35-40, 2000.

[8] V. Iyengar and K. Chakrabarty, "Precedence-based, preemptive, and power-constrained test scheduling for system-on-a-chip", accepted for publication in *Proc. VTS*, 2001.

[9] M. Sugihara, H. Date and H. Yasuura, "A novel test methodology for core-based system LSIs and a testing time minimization problem", *Proc. ITC*, pp. 465–472, 1998.

[10] A. Jas and N. A. Touba, "Test vector decompression via cyclical scan chains and its application to testing core-based design", *Proc. ITC*, pp. 458-464, 1998.

[11] A. Chandra and K. Chakrabarty, "System-on-a-chip test data compression and decompression architectures based on Golomb codes", *IEEE Trans. on CAD/ICAS*, vol. 20, March 2001 (accepted for publication).

[12] I. Hamzaoglu and J. H. Patel, "Test set compaction algorithms for combinational circuits", *Proc. Int. Conf. on CAD*, pp. 283-289, 1998.