Guided Test Generation for Isolation and Detection of Embedded Trojans in ICs

Mainak Banga, Maheshwar Chandrasekar, Lei Fang, Michael S. Hsiao
Department of ECE, Virginia Tech., Blacksburg, VA – 24061.
{banga, chandram, leifang, mhsiao}@vt.edu

ABSTRACT

Testing the genuineness of a manufactured chip is an important step in an IC product life cycle. This becomes more prominent with the outsourcing of the manufacturing process, since the manufacturer may tamper the internal circuit behavior using Trojan circuits in the original design. Traditional testing methods cannot detect these stealthy Trojans because the triggering scenario, which activates it, is unknown. Recently, approaches based on side-channel analysis have shown promising results in detecting Trojans. In this paper, we propose a novel test generation technique that aims at magnifying the disparity between side-channel signal waveforms of tampered and genuine circuits to indicate the possibility of internal tampering. Experimental results indicate that our approach could magnify the likelihood of Trojans 4 to 20 times more than existing side-channel analysis based approaches.

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

General Terms: Reliability, Security

1. INTRODUCTION

Outsourcing manufacturing process has become a trend recently, in both public and private sectors. This raises the question of "authenticity of the shipped products". It is essential to know that the oversea manufacturer did not tamper with the internals of the original design (in the manufactured chips). The part of circuit that tampers the original design behavior is usually referred to as a *Trojan* in Integrated Circuits (IC).

Trojan circuits are extraneous logic inserted in the actual design before and/or during manufacturing. These are hard-to-detect using conventional testing mechanisms. Approaches based on LFSR and Logic BIST [4, 5, 6] have been proposed to monitor the proper operation of the internal hardware. However, Trojans can be intelligently built to deter the advantages of such vigilant approaches. Destructive testing of a few chips does not guarantee the authenticity of other chips not subjected to such testing. Further, adding to the stealthy nature, Trojans can be implanted without affecting the circuit's characteristics like physical form-factor, pin numbers and die size. Trojans are generally passive for the most part of the circuit's operation [2]. However, once triggered they could result in catastrophic consequences leading to disruption of the normal functionality of the underlying design.

GLSVLSI'08, May 4–6, 2008, Orlando, Florida, USA.

Recently, in [1], the authors proposed a side-channel analysis to determine the existence of a Trojan in a given IC. To our knowledge, this is the first proposed method in the literature to determine the authenticity of third-party manufactured chips. While one may determine the genuineness of a chip by destructive testing, the non-destructive nature of power signature analysis makes this technique attractive. As the authors in [1] have identified that the power signature difference must exceed process variation to be statistically significant, some intelligent Trojan's may hide the discrepancy in signatures within process variation levels. Such Trojans become difficult to detect. Furthermore, the authors in [1] employ a (random) non-redundant set of tests. The non-discriminate nature from random test patterns may not be ideal in maximizing the discrepancy in the power signatures.

In this paper, we propose a two-step test generation technique that targets at magnifying the discrepancy between the CUT (circuit under test) and the genuine design waveforms. In the first stage, we generate intelligent test patterns such that the power signatures can be used to identify candidate regions that might be affected by a Trojan. Next, we generate new test patterns concentrating on the identified regions in order to magnify the disparity between signal waveforms. Furthermore, our approach can identify the possible regions that might contain the Trojan. Such an isolation is essential for diagnosis of the altered chip. Experimental results on ISCAS'89 circuits reveal that our approach outperforms existing methods in detecting embedded Trojans.

2. PRELIMINARIES

2.1 Side Channel Analysis

– In manufactured ICs, one can analyze parameters like electromagnetic radiation, I/O timing behavior or power profile of the circuit to assess the behavior of the overall system. Such parameters that act like a signature for the device are commonly known as the side channel signals. The method of using side channel signals to extract internal information of a device is known as the side channel analysis, and side channel analyses have been effectively used to detect the anomalies in the behavior of a circuit [7, 9].

For our approach, we compute the power profile of the genuine CUT. The total power for an IC is proportional to the operating frequency f, switching capacitance C, and supply voltage V, shown in the following expression [3]:

$$P \propto CV^2 f$$

As the overall power consumption will reduce if the circuit is operated at a low frequency, simple Trojans could be more easily detected because the power consumed by Trojan will make up a greater portion in the total consumed power. This was illustrated in [1] by an experiment in which a simple Trojan could not be detected when the circuit was operated at 100 MHz, whereas it was detected at 500 KHz.

2.2 Hamming Distance – For two states in a circuit, the Hamming distance between them is defined as the number of bit differences between the states.

In our approach, we try to minimize the overall switching activity. Since the power consumed in the circuit is directly proportional to the amount of switching activity occurring in it [8, 10], by minimizing the switching activity, we actually try to minimize the total power consumption. We use the following terms for our subsequent discussion:

Combinational Trojan: A combinational circuit that becomes active when a specific condition arises in the internal signals and/or circuit flip-flops or a portion of it.

Sequential Trojan: A finite state machine (FSM) that monitors a portion of the internal circuit signals and triggers the output upon the occurrence of specific sequence(s).

2.3 Power Profile – The total power consumed in the circuit over a set of vectors constitute the power profile of the circuit for that vector set and the individual power value for any particular vector-pair is called the power number for that vector-pair. Frequently, the power numbers may be estimated by parameters such as the circuit's switching activity.

3. OUR APPROACH

Our approach consists of two steps. In the first step, the test set quickly and intelligently sweeps through the state space in a controlled manner and generates activity within subsets of flip-flops while keeping the activity of the rest of the circuit low. After analyzing this power profile, we identify possible subsets of state signals that may feed the Trojan in the circuit. In the second step, we focus on those regions identified in the first step and generate a new test suite to further increase the relative difference in the power profiles between the actual circuit and the Trojan counterpart. In this step, if we observe a sustained increased activity over the expected behavior, it clearly indicates anomalous behavior that strongly suggests the presence of a Trojan. We call these two steps as "Circuit Partitioning" and "Activity Magnification" respectively, and they are detailed below.

3.1 Stage1: Circuit Partitioning

Trojans usually constitute a tiny fraction of the total chip area. It is intuitive that during normal functional operation, the activity in the overall circuit could be several orders of magnitude greater than the activity of the Trojan. Hence, the relative increase in the circuit activity due to the presence of the Trojan may not be above the process variation, and consequently it might be difficult to make any inference about its presence. Therefore, in order to detect a Trojan circuit, we need to increase the activity within the Trojan portion of the circuit while simultaneously minimizing the activity for the rest of the circuit. We note that we should not decrease the power so low such that the CUT enters some sleep mode. If the Trojan also enters the sleep mode, we will not be able to observe discrepancies in the power signatures.

Since we cannot predict the location of the Trojan in the circuit, we use a "divide and conquer" approach as an attempt to isolate it. In general, we can broadly classify the flip-flops in a circuit into different groups depending on the functionality with which they are associated. Trojans being intelligent monitors, their triggering condition is likely to be associated with one or more such functional groups. Hence, it is better to focus on a smaller portion of the state space than the complete set of flip-flops considered together.

Consider a circuit with n flip-flops. Given a subset of flip-flops, G, the signals and gates that lie in the fanout cone of G defines a region of interest. Our algorithm partitions the circuit into small regions based on structural connectivity. At any point during test generation, we try to increase the activity in the corresponding region of interest while keeping the rest of the circuit at low activity. To do so, we maximize the Hamming distance between any two successive states in the subset G, while simultaneously minimizing the Hamming distance for the rest of the state variables. This is important because we do not want the power from the non-Trojan part of the circuit to drown out the power from the Trojan. By minimizing the Hamming distance for the rest of the flip-flops, the signals in the fanout cone of these flip-flops undergo little activity thereby reducing the overall circuit activity. We calculate the per flip-flop increase in the Hamming distance for the group G as well as for all other flip-flops that are not in G. The difference between these two quantities serves as the selection parameter for an appropriate input vector from a list of available vectors.

Let S be the entire set of flip-flops in the circuit. Again, let G be a group of flip-flops for which we are maximizing the Hamming distance. Let d be the Hamming distance for the flip-flops in G and d' be the Hamming distance for the rest of the flip-flops. Then, we define our objective function F as the following:

$$F = max\ (d/g - d'/g') \quad \ldots (1)$$

where g is the number of flip-flops in the group G and g' is the number of flip-flops in the rest of the circuit apart from those in G.

We simply generate k random input vectors and select the best vector-pair from within it. We repeat this until we have obtained enough vectors. We note that a large value of k ensures that we get a good vector-pair. On the other hand, k should be small enough so that we do not incur a major runtime penalty. In our experiments, we limit k to be less than 20 for each subset, and we repeat the process for all the subsets of flip-flops in the circuit.

3.2 Stage 2: Activity Magnification

Based on the comparison of the relative difference in the power profiles for the genuine and Trojan circuits using the vector sequence generated in Stage 1, we identify the regions (set of flip-flops) that exhibit increased relative activity. In this stage, we generate more vectors for the specific region(s) marked as possible regions containing the Trojan using the same test generation approach as discussed in Stage 1. Results show that our method significantly magnifies the relative activity difference between the Trojan and the genuine circuit.

The flow of our overall approach is represented in the flowchart shown in Figure 1.

4. TROJAN SET-UP

In our work, all the Trojans used are less than 1% of the gate count of the original circuit (which equivalently translates to chip area). We have ensured that once triggered, the Trojan affects one or more parts of the circuit impairing the normal internal functionality. Furthermore, we note that the Trojans are difficult-to detect by confirming that the output generated by simulating the vectors exactly match for both the genuine and affected circuits. Otherwise, if we could easily detect the Trojan at a primary output, the side channel analysis would not have been required in the first place.

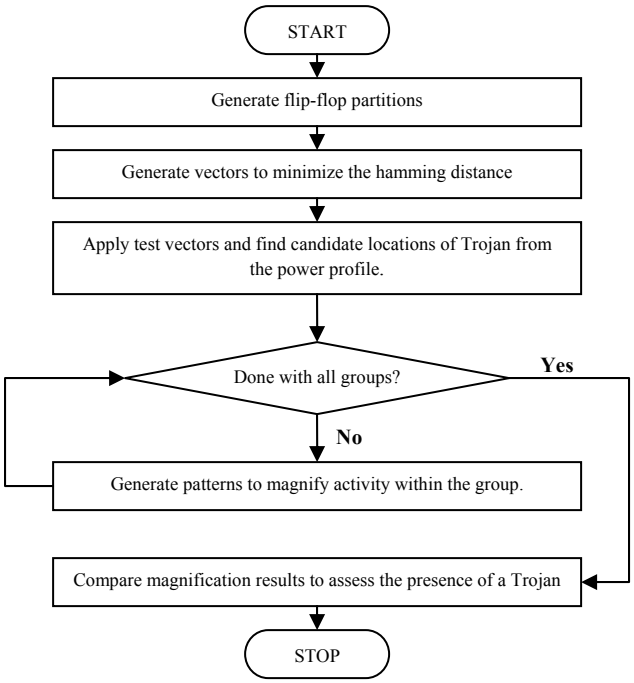


Figure 1: Overall flow of the Trojan identification Process

A standalone embedded Trojan circuitry is shown in Figure 2. For a given CUT, the malicious circuitry is obtained by connecting the inputs of the Trojan to the appropriate flip-flops and the output of the Trojan to appropriate gate(s) of the genuine circuit. A state-transition analysis helps us to observe that for most of the operation cycle, the output of the Trojan circuit is logic 0. In fact, in our experiments, we have generated vector sets consisting of 200 vectors on an average and it contains only one such sequence that triggers the Trojan (sets output to 1). The stealthy nature of a Trojan is modeled by ensuring that even if the Trojan is triggered, the discrepancy does not reach the output(s).

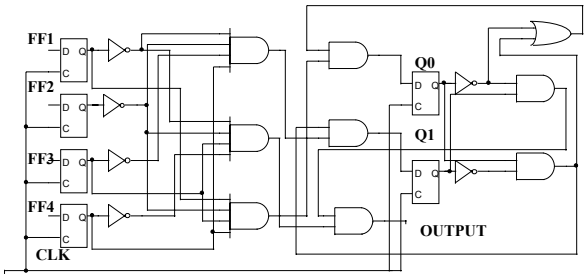


Figure 2: A sample Trojan circuit

In our approach, we start with an initial reset state of 0s and generate the vectors based on the heuristic defined by Equation 1. For groups with 5 flip-flops we generate 20 patterns per group while for groups with 10 or 20 flip-flops we generate 10 patterns per group. We compare the power profiles of the Trojan and genuine circuits, and our results show that we can identify regions that show relatively high activity as compared with the random power profile. Note that since random vectors do not distinguish areas in the entire circuit, no specific information about the region of the Trojan can be deduced from the power profile. Next, we probe into these identified regions in Stage 2. Our analysis assesses the extent of extraneous activity more elaborately and confirms whether process variation alone can account for the discrepancies observed.

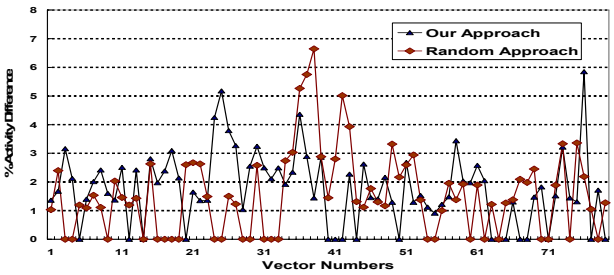
5. EXPERIMENTAL RESULTS

5.1 Stage 1: Circuit Partitioning Results

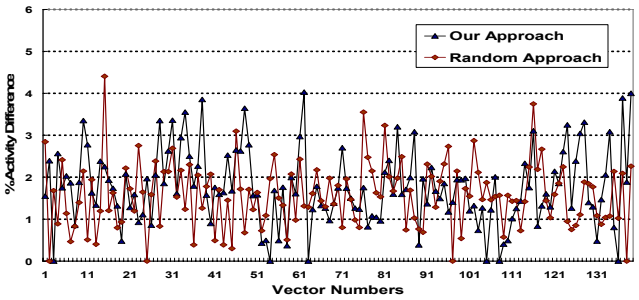
In each of the graphs the X-axis denotes the index for vector numbers, while the Y-axis denotes the percentage difference in activity between the Trojan circuit and the genuine circuit.

For s1196, the black (triangular legend) curve in Graph 1 shows that the percentage activity difference between actual circuit and the Trojan circuit is amplified in the regions covered by our generated vectors 15 to 20 and between vectors 23 to 34. The Trojan is indeed connected to the second group excited by vector numbers 21-40. The difference in the magnification obtained by our approach as compared to the random clearly separates out the second region for further magnification.

In Graph 2 for circuit s3330, we generated 10 vectors per group. We can separate out regions corresponding to flip-flop groups 3 (which is covered by vectors 21-30), 4, 5, 7, 9, 12 and 13 as the portions with distinct increase in the percentage circuit activity. In our experiment, we have associated the Trojan with a portion of the flip-flops in group 5 which we could isolate as a target region for further analysis in Stage 2.

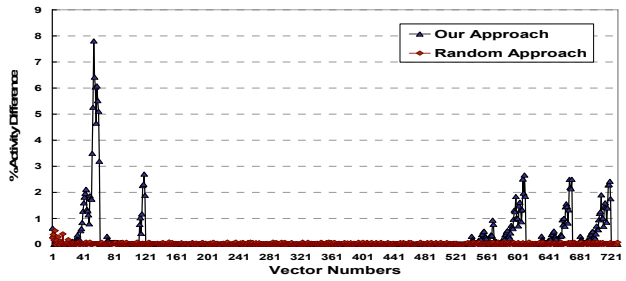


Graph 1: The relative increase in Trojan circuit activity by our approach vs. the random approach for s1196



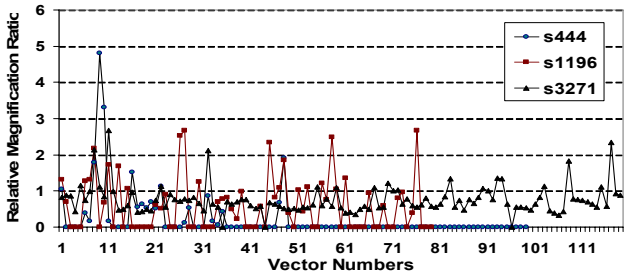
Graph 2: The relative increase in Trojan circuit activity by our approach vs. the random approach for s3330

Graph 3 reveals that the random method hardly shows any difference in the percentage activity between the genuine and the Trojan circuits. However, our approach separates out distinct regions viz. flip-flop groups 5, 6, 12, 61, 67, 71 and 72 where the extraneous activity in the Trojan is high enough to produce a difference as high as 8% from the actual circuit. This is the graph for circuit s38584 for which the Trojan is embedded into the 72nd group represented by the vectors 711 to 720.

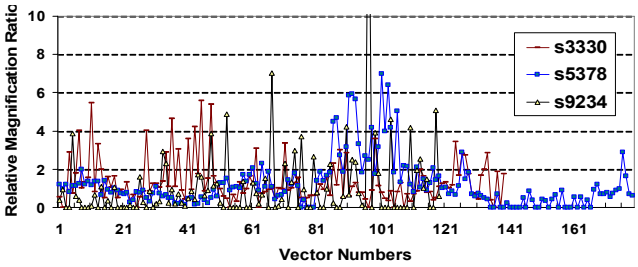


Graph 3: The relative increase in Trojan circuit activity by our approach vs. the random approach for s38584

Graphs 4 and 5 give magnification ratio of the Trojan circuit activity using our method as compared to that of the random method. We observe that our method magnifies the Trojan to actual circuit activity by 4 to 20 times in the portions which are identified as candidate regions. These two graphs show that our stage 1 can consistently locate the candidate Trojan regions.



Graph 4: Ratio of relative magnification of Trojan circuit activity over the actual circuit activity for different circuits

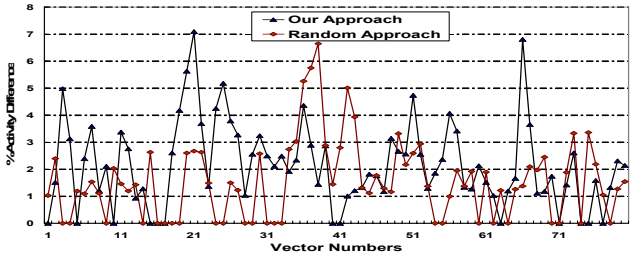


Graph 5: Ratio of relative magnification of Trojan circuit activity over the actual circuit activity for different circuits

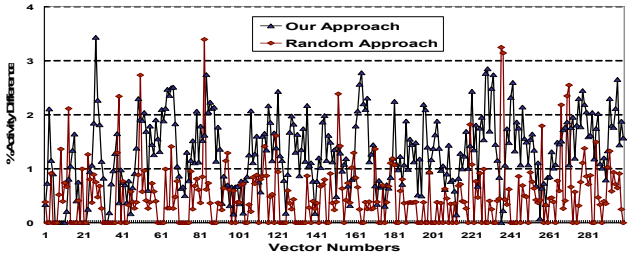
5.2 Stage 2: Activity Magnification

Our attempt to magnify the activity for target groups in circuit s1196 shows that when we zero in on those regions most responsible for the Trojan, the magnification of the power dissipation ratios is significant when compared to the random vectors. Graph 6 shows these results. An important observation is that, at times, we are able to achieve a magnification in the activity of the Trojan from the actual circuit in excess of 6% (which is normally greater than the process variation) and this trend is not observed in the graph obtained at the first stage.

s15850 shows one of the best results for the activity magnification step, shown in Graph 7. In this circuit, the Trojan is connected to 27th flip-flop group and when we attempted to zero-in on the power numbers for the 27th group, it clearly indicated that the targeted group produces noticeable extraneous activity as compared to the random vectors.



Graph 6: Activity Magnification for s1196, (group 2) between our approach vs. random approach



Graph 7: Activity Magnification for s15850, between our approach vs. random approach

6. CONCLUSIONS

We have presented a two-stage approach to generate a set of effective test cases that is able to detect the presence of a Trojan in a given design. Experiments showed that our method is able to provide a 4 to 20 times magnification in the circuit activity for the circuit with a Trojan over a genuine circuit. Moreover, in circuits like s38584 our method points the target areas distinctly where the conventional random patterns fail to make any distinction.

REFERENCES

[1] D. Agarwal, S. Baktir, D. Karakoy, P. Rohatgi, B. Sunar, "Trojan Detection using IC Fingerprinting", IBM Research Report, 2006.

[2] K. Nowaka, G. Carpenter, F. Gebara, J. Schaub, D. Agarwal, P. Rohatgi, W. E. Hall, S. Baktir, D. Karakoyunlu, B. Sunar, "IC Fingerprinting and Stable IS Sensors for Enhanced IC Trust, 2006.

[3] S. Pilli, S. S. Sapatnekar, "Power estimation considering statistical IC parametric variations"; ISCAS 1997, pp. 1524 – 1527, vol.3.

[4] C. Fagot, O. Gascuel, P. Girard, C. Landrault, "On Calculating Efficient LFSR Seeds for Built-In Self Test"; Proc. Of European Test Workshop, 1999, pp 7-14.

[5] G. Hetherington, T. Fryars, N. Tamarapalli, M. Kassab, A. Hassan, J. Rajski, "Logic BIST for large industrial designs: real issues and case studies"; ITC, 1999, pp. 358-367.

[6] W.-T. Cheng; M. Sharma; T. Rinderknecht, C. Hill, "Signature Based Diagnosis for Logic BIST"; ITC 2006, Oct. 2006, pp. 1 – 9.

[7] L. J. Kohout, A. Yasinsac, E. McDuffie, "Activity profiles for intrusion detection" ; Fuzzy Information Processing Society, 2002. pp. 463 – 468.

[8] W. Li; S. M. Reddy, I. Pomeranz, "On reducing peak current and power during test"; Proc. IEEE computer society annual symposium, 2005, pp. 156 – 161.

[9] D. Agarwal. et al, "The EM side-channel(s)" CHES 2002, v 2523 Lecture Notes on Computer Science, Springer-Verlag, pp. 29-45, 2002.

[10] F. N. Najm, "Transition density: a new measure of activity in digital circuits;", IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol 12, Issue 2, Feb. 1993 pp. 310 – 323.