Sensitivity Analysis to Hardware Trojans using Power Supply Transient Signals

Reza Rad
reza2@umbc.edu
University of Maryland, B.C.

Jim Plusquellic
plusquel@umbc.edu
University of New Mexico

Mohammad Tehranipoor
tehrani@engr.uconn.edu
University of Connecticut

ABSTACT

Trust in reference to integrated circuits addresses the concern that the design and/or fabrication of the IC may be purposely altered by an adversary. The insertion of a hardware Trojan involves a deliberate and malicious change to an IC that adds or removes functionality or reduces its reliability. Trojans are designed to disable and/or destroy the IC at some future time or they may serve to leak confidential information covertly to the adversary. Trojans are cleverly hidden by the adversary to make it extremely difficult for chip validation processes, such as manufacturing test, to accidentally discover them. This paper investigates a power supply transient signal analysis method for detecting Trojans that is based on the analysis of multiple power port signals. In particular, we focus on determining the smallest detectable Trojan in a set of process simulation models that characterize a TSMC 0.18 um process.

1.0 Introduction

The globalization of the integrated circuit industry in combination with the dramatic increases in the complexity of ICs have raised new concerns regarding the integrity of ICs [1][2]. The threat is the malicious modification to the function of a IC or the inclusion of additional circuitry designed to enable an adversary to destroy, disable or remotely control the IC through a back door at a time of his or her choosing. Logic-based testing techniques designed to uncover the presence of Trojans are not likely to be effective against even the simplest Trojan hiding techniques. Techniques that relay on physical inspection and destructive reverse engineering are difficult and costly.

Parametric testing techniques such as those based on the analysis of power supply signals are better suited for Trojan detection but require modifications for several reasons. First, the adversary can configure the Trojan to have a minimal impact on the IC's nominal quiescent (I_{DDQ}) or transient (I_{DDT}) power supply current. Therefore, conventional testing methods that measure global, chip-wide, behavior of I_{DDQ} or I_{DDT}, are ineffective because of very small Trojan-signal-to-background-current ratios that are present in multi-million transistor chips. Second, process variations are increasing significantly in advanced technology nodes, making it more difficult to differentiate between signal variations introduced by process anomalies and those introduced by Trojans. Third, given process noise, statistical detection techniques need to be applied. Unlike existing 'data-driven' power supply analysis techniques for defects, statistical thresholds for Trojans need to be developed from simulation models. This is true because all or a large fraction of the ICs may contain Trojans and cannot be used in a data-driven methodology to define the statistical thresholds.

We describe a hardware Trojan detection method that addresses these issues. The method analyzes an IC's supply current measured from multiple supply ports to deal with the small Trojan-signal-to-background-current ratios. Simple calibration circuits and procedures are used to reduce the adverse impact of process variations on Trojan sensitivity. The calibration technique transforms the measured currents for each IC to match those produced from a golden, Trojan-free simulation model. This transformation process greatly amplifies Trojan and defect anomalies. A Trojan is easily distinguished from a random defect by observing patterns in the detection process across multiple ICs.

In this paper, we describe our Trojan detection method and evaluate its detection capabilities using simulation experiments. Ten different layouts containing an ISCAS '85 benchmark circuit are constructed in a TSMC 0.18 um six metal layer process [3][4]. One of the layouts is Trojan-free. In the nine remaining layouts, extra gates are added to model the Trojan, with each layout containing one more extra gate than the previous layout. Simulation models are extracted from the layouts and simulation data is analyzed to determine when it is possible to detect the Trojan. The results of our sensitivity analysis indicate that it is possible to reliably detect 'un-activated' Trojans implemented using as few as four standard cell gates.

2.0 Background

There are a wide variety of hardware security issues that have been recently addressed. References [5-8] provide a few examples from a growing body of research in this area. However, hardware Trojans are a new concern and consequently, the literature has few published works on the topic.

The use of power supply transient signals, also called side-channel signals, for the detection of Trojan signals is proposed in [9]. The authors analyze a single global transient in their proposed method and therefore, their method does not scale with larger ICs. Moreover, process variations have a significant impact on the measured transients and therefore, a method for calibrating for them is essential.

We have proposed several power supply detection and localization methods in previous work for detecting manufacturing defects [10][11][12]. In this paper, we propose a tech-

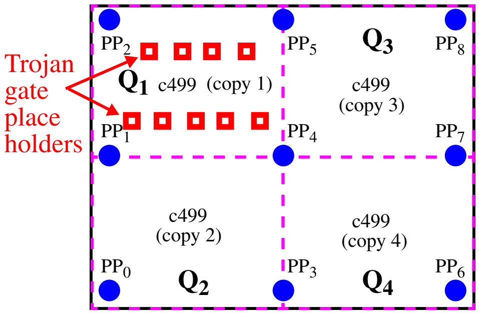

Figure 1. Architecture of simulation model

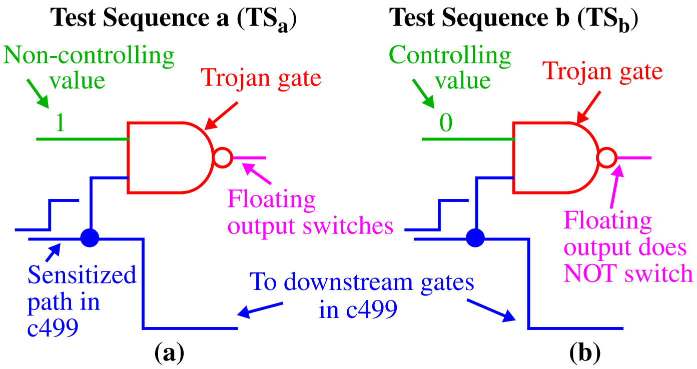

Figure 2. Trojan gate sensitization criteria under test sequences

nique based on the analysis of power supply transient signals and develop a statistical *prediction ellipse* technique for detecting Trojans.

3.0 Experiment Setup

Figure 1 shows a top level view of the design used in the simulation experiments. It consists of four 'quads' labeled Q_1 through Q_4. A copy of the c499 ISCAS '85 benchmark circuit is inserted into each quad. The layout of the c499 in Quad Q_1 is modified to include nine empty rectangles. These rectangles are 'holes' in the layout and correspond to the size of a typical standard cell. Trojans are modeled by inserting a gate into one or more of these rectangles, as explained below.

The design in Figure 1 was constructed using the technology rules for the TSMC 0.18 um process [4]. The power grid is routed in a standard mesh configuration over all six metal layers available in the process. Nine power ports, labeled PP_0 through PP_9, connect to the power grid in the top most metal layer. The ground grid is interleaved with the power grid (not shown) and is configured in a similar fashion.

Ten different layouts of the design were constructed. In the Trojan-free version, all nine of the rectangles are empty. In the first Trojan layout, a standard cell gate is inserted into one of the rectangles and is connected to nodes in the surrounding neighborhood. This process is repeated for each of the remaining (eight) Trojan layouts, with one additional standard cell gate added in each successive layout. The Trojan models are referred to as T_1 (one gate version) through T_9 (nine gate version). The position of the rectangles are kept constant in all ten layouts to minimize the differences.

A set of fifteen simulation models are extracted from the Trojan-free layout using published process parameters for the TSMC 0.18 um process. Ten of the simulation models are used to define statistical limits as described in Section 4.2. The remaining five are used as control samples to evaluate the false alarm rate of our Trojan detection method. False alarms occur when a Trojan is detected in one of the Trojan-free simulation models that are designated as control. For each of the nine Trojan designs, ten simulation models are extracted using the same process models as those used to create the Trojan-free (non-control) versions.

Two test sequences are used as a stimulus to evaluate Trojan sensitivity. Both test sequences propagated signals along the same paths in the Q_1 copy of the c499 shown in Figure 1. The difference in the stimuli is related to the off-path inputs of the inserted Trojan gates. Figure 2 shows the two scenarios, labeled (a) and (b), using an NAND gate to represent the Trojan. One of the inputs of the NAND is connected to a sensitized path in the c499. The other input is held constant at one of two values. In (a), a non-controlling value is placed on the off-path input. Therefore, when the test sequence is applied to the PIs of the c499, the propagating signal causes the NAND gate's output to switch (and consume power). We refer to this test sequence as TS_a. In (b), a controlling value is placed on the off-path input preventing the NAND gate from switching. Therefore, only the capacitive loading of the on-path input can affect the power consumption. We refer to this test sequence as TS_b.

Note that the output of the Trojan is not connected in our experiments. For an actual Trojan instance, this would not be the case. Therefore, our Trojan models minimize the impact of the Trojan on power consumption, better suiting the objective of our sensitivity analysis. The configurations shown in Figure 2 are replicated for Trojan models that incorporate more than one gate, i.e., T_2 through T_9.

In addition to the two test stimuli, our method requires the application of a set of calibration tests. These are briefly outlined in the next section. A full description of the calibration process and it's impact on detection sensitivity is given in reference [13].

4.0 Trojan Detection using Power Supply Transient Signals (I_{DDT})

The power supply transient analysis (I_{DDT}) technique that we propose analyzes *local* I_{DDT} measurements obtained from the multiple, individual *power ports* on the chip. The I_{DDT} signals are measured from each of the power ports as a test sequence is applied to the inputs of the core logic. The I_{DDT}s of neighboring power ports, e.g. PP_0 and PP_1 in Figure 1, are compared to identify anomalies introduced by the presence of a

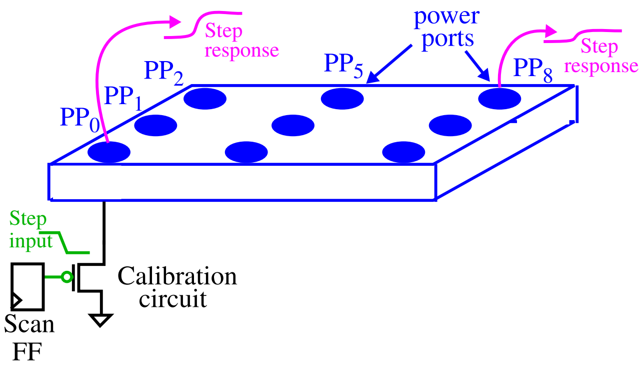

Figure 3. Calibration circuit step responses (left) and their integrated derivatives.

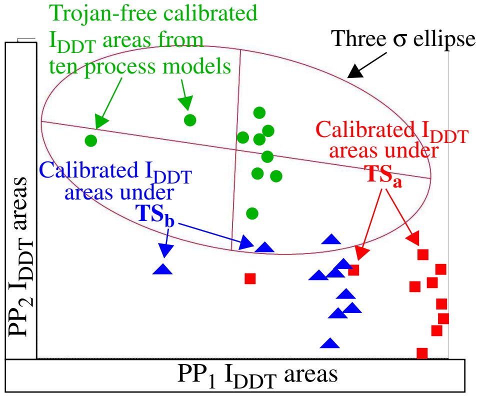

Figure 4. Scatterplot of Trojan-free I_{DDT}s and passive and active 3-gate Trojan (T_3) from 10 process models for power port pairing PP_1 and PP_2.

Trojan circuit. We describe a statistical method to identify the anomalies in Section 4.2. In order to reduce the adverse effects of process variations, we first apply a set of calibration tests as described below.

4.1 Calibration

Calibration is used to deal with process variations that occur in the chip's core logic, power grid and off-chip connections to the power ports. Calibration is carried out on each IC using a set of inserted calibration circuits. Figure 3 shows a calibration circuit represented by a p-channel transistor whose gate is connected to the output of a scan flip-flop (FF). The source of the p-channel is connected to the power grid in metal 1 directly underneath the power port, e.g., PP_0 in the figure. Although not shown in Figure 3, the same arrangement would be implemented for each of the other power ports.

A calibration test is carried out by configuring the scan chain to deliver a step input to the gate of the p-channel transistor. This can be implemented by scanning a 0 through the scan chain with all other elements initialized to 1. The step input enables the p-channel transistor and effectively creates a short between V_{DD} and GND. The step input response is then measured at each of the power ports. Example waveforms are shown for PP_0 and PP_8 in the figure. The *impulse responses* are obtained from the step responses by taking the derivative. The areas under the derivative waveforms are computed and inserted into a calibration matrix. The nine power ports and nine calibration tests produce a set of eighty-one areas that define the matrix. This process is repeated using a 'golden' simulation model. The two matrices define a transformation matrix that is subsequently used to calibrate the waveforms measured under the Trojan tests. The calibration process is described in detail in reference [13].

4.2 Scatterplot Analysis for Trojan Detection

We propose a robust statistical analysis technique to detect transient signal anomalies introduced by Trojans. As indicated in Section 3.0, we apply two test sequences to each of the one hundred simulation models and measure the I_{DDT} areas produced on each of the nine power ports. The areas from the ten Trojan-free simulation models are first calibrated to reduce the adverse impact of process variations. The calibrated areas from pairs of neighboring power ports are then plotted in two-dimensional scatterplots. The mean and variation of the data points in each scatterplot are used to derive the statistical limits implemented as an enclosing ellipse. The region enclosed by the ellipse defines the space in which the data points from Trojan-free ICs are expected to fall. Data points that fall outside the limits are deemed to belong to an IC with a Trojan.

As an example, Figure 4 shows the scatterplot of calibrated I_{DDT} areas for supply port pairing PP_1 (x-axis) and PP_2 (y-axis). The circles represent the data points from the ten Trojan-free simulation models under either of the test sequences. A three sigma *prediction ellipse* is derived from these points and defines the Trojan-free space. The elliptical bound is computed from the eigen values of the Trojan-free covariance matrix and a three σ X^2 (chi-square) distribution statistic. The square data points are obtained from the ten T_3 (3-gate Trojan) process models under the test sequence TS_a (see Figure 2) while the triangular data points are those obtained under the test sequence TS_b. Since this power port pairing is adjacent to the quad with the inserted Trojan, all except two of the T_3 data points under either test sequence are detected as outliers. Here, an outlier is defined as a data point that falls outside of the three σ ellipse.

The process of determining whether a given Trojan under a specific process model is detected is based on outlier analysis of scatterplot data. For each Trojan model, twelve scatterplots are analyzed, one for each adjacent power port pairing as shown in Figure 1, e.g., PP_0-PP_1, PP_1-PP_2, PP_0-PP_3, PP_1-PP_4, etc. If any one of the twelve Trojan data point falls outside the prediction ellipse limits across the twelve scatterplots, the Trojan is identified as detected. For example, the analysis of the scatterplot given in Figure 4 yields a positive detection for eighteen of the twenty 3-gate Trojan models. The detection algorithm would continue to evaluate the eleven remaining scatterplots (not shown) and count the number of detections for

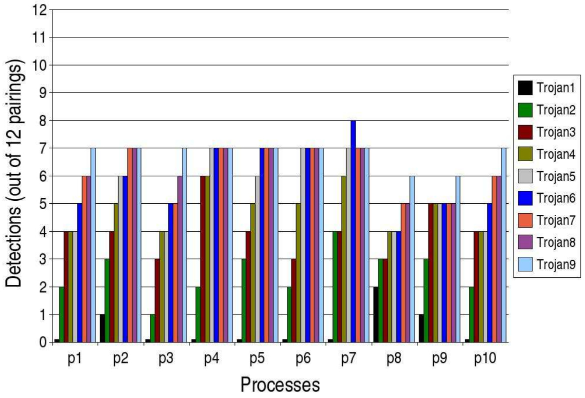

Figure 5. Detection results of ninety Trojans under test sequence (TS_a).

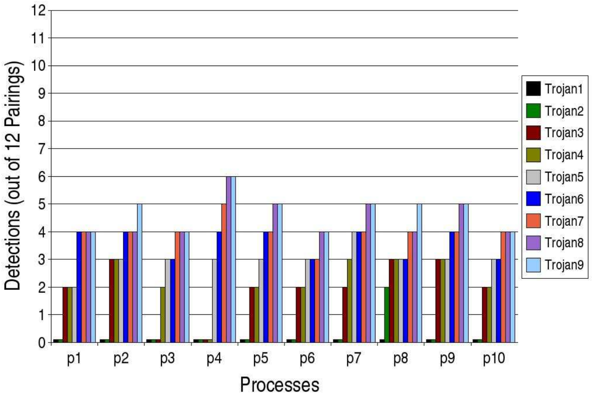

Figure 6. Detection results of ninety Trojans under test sequence (TS_b)

each Trojan model.

5.0 Simulation Results

As indicated in Section 3.0, we constructed fifteen Trojan-free simulation models. We use ten of those models to define the three sigma prediction ellipses for each of the twelve scatterplots. The remaining five Trojan-free models are used as control samples. The data points from the control samples all fall within the ellipses and therefore, in these experiments, there are no false alarms to report.

Figures 5 and 6 give the results for the ninety Trojan experiments under test sequences TS_a and TS_b respectively. The x-axis lists the ninety experiments grouped in ten clusters. The labels *p1* through *p10* label the process model associated with each of the clusters. Each cluster contains nine bars, one for each of the nine Trojans. The height of the bar indicates the number of data points that fell outside of the limits across the twelve scatterplots. Larger bars indicate more detections and provide higher confidence in the detection decision.

As discussed in Section 3.0, test sequence TS_a assigns non-controlling values to the off-path inputs of the Trojan gate(s). This enables the Trojan gate(s) to switch as the on-path input (the input connected to the sensitized path in the c499) toggles. Therefore, the number of detections under TS_a is expected to be larger than the number under TS_b. This trend is clearly visible by comparing the height of the bars in Figures 5 and 6. In fact, it is always true that the number of detections for a Trojan under any given process model in Figure 5 is larger than the corresponding number in Figure 6. A second important trend that is clear in the results is that the number of detections increases as the number of Trojan gates increases. For example, T_1 (one-gate Trojan) is not detected except under three process models in Figure 5 (*p2*, *p8* and *p9*), while T_2 is detected in all process models and is detected more often. This is expected since T_2 is implemented using the gate from T_1 plus one additional gate. A similar trend is observable for the T_3 through T_9 in both sets of results.

The goal of these experiments is to determine the sensitivity of our methods to Trojans. The results in Figures 5 and 6 indicate that we can reliably detect Trojans as small as two gates if both gates switch in response to a test sequence. This number increases to four gates for Trojans that do not switch but connect to a sensitized path. Note that it may be possible to identify the presence of one gate Trojans under TS_a and two gate Trojans under TS_b because the number of detections is non-zero for some process models. In other words, detecting a Trojan in every corner of the process space is not required. If anomalies are detected in even a small fraction of the ICs tested, this should prompt the application of a more thorough inspection process.

6.0 Conclusions

A test strategy that detects anomalies introduced by the Trojan in the power port currents is proposed. A calibration technique is proposed to deal with the adverse effects of process variations on Trojan resolution. A statistical analysis procedure is defined that enables the detection of Trojan anomalies in the power supply transient currents of an IC. Our simulation results demonstrate that it is possible to reliably detect Trojans implemented with as few as four gates and that activation of the Trojan is not necessary. We are currently investigating our method on a large microprocessor architecture to determine the scalability of our technique.

7.0 Acknowledgements

The work of Reza Rad and Jim Plusquellic was supported in part by NSF grant CNS-0716559. The work of Mohammad Tehranipoor was supported in part by NSF grant CNS-0716535.

8.0 References

[1] http://www.acq.osd.mil/dsb/reports/2005-02-HPMS_Report_Final.pdf
[2] http://www.darpa.mil/mto/solicitations/baa07-24/index.html
[3] http://www.fm.vslib.cz/~kes/asic/iscas/
[4] http://www.mosis.com/Technical/Testdata/tsmc-018-prm.html

[5] B. Yang, K. Wu, and R. Karri, “Scan Based Side Channel Attack on Dedicated Hardware Implementations of Data Encryption Standard,” in *Proc. of the IEEE Int. Test Conf. (ITC)*, pp. 339.344, 2004.

[6] S. Ravi, A. Raghunathan, and S. Chakradhar, “Tamper Resistance Mechanisms for Secure Embedded Systems,” in *Proc. of the 17th Intl. Conf. on VLSI Design*, pp. 605.611, 2004.

[7] P. Kocher, R. Lee, G. McGraw, A. Raghunathan, and S. Ravi, “Security as a New Dimension in Embedded System Design,” in *Proc. of the 41st Annual Conference on Design Automation*, pp. 753.760, June 2004.

[8] K. Tiri and I. Verbauwhede, “A VLSI Design Flow for Secure Side-Channel Attack Resistant ICs,” in *Proc. of Design, Automation and Test in Europe*, pp. 58.63, Mar. 2005.

[9] D. Agrawal, S. Baktir, D. Karakoyunlu, P. Rohatgi, B. Sunar, “Trojan Detection using IC Fingerprinting”, Symposium on Security and Privacy, 2007, pp. 296 - 310.

[10] A. Germida, Z. Yan, J. Plusquellic and F. Muradali, “Defect Detection using Power Supply Transient Signal Analysis,” in Proc. *Int. Test Conf.*, pp. 67-76, Sept. 1999.

[11] J. Plusquellic, D. Acharyya, A. Singh, M. Tehranipoor and C. Patel, “Quiescent Signal Analysis: a Multiple Supply Pad I_{DDQ} Method,” IEEE Design and Test of Computers, vol. 23, no. 4, pp. 278-293, 2006.

[12] D. Acharyya and J. Plusquellic, “Hardware Results Demonstrating Defect Detection Using Power Supply Signal Measurements,” in Proc. *VLSI Test Symp. (VTS’05)*, pp. 433-438, 2005.

[13] Reza M. Rad, Xiaoxiao Wang, Mohammad Tehranipoor, Jim Plusquellic, “Taxonomy of Trojans and Methods of Detection for IC Trust”, submitted to ICCAD 2008.