A study on the effectiveness of trojan detection techniques using a red team blue team approach

X. Zhang, K. Xiao, M. Tehranipoor
Department of Electrical and Computer Engineering
University of Connecticut
Email: {xuz09001, kanxiao, tehrani}@engr.uconn.edu

J. Rajendran and R. Karri
Department of Electrical and Computer Engineering
Polytechnic Institute of New York University
Email: jrajen01@students.poly.edu, rkarri@poly.edu

***Abstract*—As part of the Embedded Systems Challenge, we assess the effectiveness of Trojan detection techniques. The red team inserted different types of Trojans – combinational, sequential, reliability degrading, and performance degrading – into selected variants of a target design; the other variants are Trojan-free. The blue team has to correctly classify the Trojan-free and Trojan-infected variants. Seven different teams from six different universities performed the blue team activity using different types of Trojan-detection techniques, namely activation-based detection, and power- and delay-based side-channels.**

I. Introduction

Globalization of the integrated circuit (IC) design flow is creating opportunities for rogue elements within the supply chain to corrupt the IC design [1]. It is possible for an attacker to insert malicious circuitry (called as *hardware Trojans*) into the hardware. To establish trust during fabrication, using trusted foundries for fabrication have been proposed [2]. However, they are not economically feasible and go against the globalization trend.

In an alternate approach to establish trust, researchers have developed different types of Trojan detection methods. These methods make it difficult for an attacker to insert Trojans. Such techniques include: (*i*) Side-channel fingerprinting – Trojan-inserted in an IC is identified by comparing its side-channel characteristics (power consumption [3], delay [4], or a combination of them [5]) with that of the golden model (Trojan-free IC). (*ii*) Hardware Trojans activation – Trojans are activated by increased switching activity within the circuit and exposing their malicious behavior [6]. (*iii*) On-chip monitors – Important signals within a circuit are monitored. When there is an unexpected behavior, the IC will be shut down or reset [7], [8]. (*iv*) Trojan modeling – A database for all existing hardware Trojan is constructed. Given a target design, the presence of Trojans from this database is checked.

To analyze the effectiveness of the Trojan detection methods, we performed a red team blue team assessment through the annual Embedded Systems Challenge competition [9], [10]. This competition involved three activities:

- **Trojan insertion – Red team activity:** The red team activity emulates the attacks launched either by an insider (malicious designer in the design team) or by a rogue element in the manufacturing phases by inserting different types of Trojans in a target design.
- **Trojan detection – Blue team activity:** The blue team then applies Trojan detection techniques of their choice and evaluates if the design was compromised. The results are then passed on to the red team.
- **Evaluation – Red & blue teams activity:** The red team analyzes the outcome of the blue team activity. If a Trojan is not detected by a detection technique, then that technique is considered weak. Then, the red team and blue team together analyze techniques to improve such weak defenses.

The paper is organized as follows: Section 2 describes the different types of Trojans inserted by the red team. Section 3 analyzes the detection techniques employed by the blue team and the outcome of the blue team. Section 4 concludes the paper.

II. Red team activity: Trojan insertion

To evaluate the effectiveness of Trojan detection techniques, red team has to perform the following tasks: 1) select target designs, 2) select target platform, and 3) design and insert Trojans.

A. Target designs and target platform

Two designs, Design A and Design B, are used in this competition. These designs are sufficiently big for the blue team to exercise different Trojan detection techniques and small enough to fit into the target platform. Design A is 2500-gate c6288, a ISCAS'85 benchmark circuit. Design B is s9234, a ISCAS'89 benchmark circuit. It has about 5000 combinational gates and 200 flip-flops. These designs were used to test the effectiveness of different Trojan detection mechanisms in [3], [4], [5].

The Xilinx Spartan3E-100 FPGA [11] is the target platform. This board has several features that can be used by a Trojan detection technique. For instance, it has support for JTAG to send in test patterns. This board has been used in previous years' Embedded System Challenges [9]. The number of input and output ports of the designs exceeded the board's capacity. Hence, we multiplexed them. Design A uses 101 look-up-tables (LUTs) and Design B uses 542 LUTs.

B. Hardware Trojan design and insertion

The red team designed and inserted six Trojans (Trojan 1 to 6) in Design A and two Trojans (Trojan 7 and 8) in Design B. These Trojans are described below:

978-1-4673-5543-8/13/$31.00 ©2013 IEEE

(*i*) *Trojan 1*: This is an input triggered Trojan. Upon activation this creates wrong outputs.

(*ii*) *Trojan 2*: This is an input triggered Trojan triggered by input vector $8'b10010110$. Upon activation, the Trojan will change the output from $8'b01000001$ to $8'b01001001$.

Trojan 1 and Trojan 2 are combinational Trojans and hence do not hold any state information. The blue team can explore all possible input vectors to activate them and observe their malicious behavior. Activation of a Trojan is relatively difficult when the Trojan consists of sequential elements. To analyze the ability of Trojan detection techniques against such Trojans, the red team designed Trojans 3 and 4.

(*iii*) *Trojan 3* is also an input triggered Trojan. It requires a sequence of input vectors $8'hCD \rightarrow 8'hCD \rightarrow 8'hCD \rightarrow 8'hCD$ to get activated. The Trojan activation probability based on this sequence is $\frac{1}{256^4}$. Upon activation, the first output bit is forced to logic 1.

(*iv*) *Trojan 4* is triggered by a sequence of input vectors "$8'hB8 \rightarrow 8'h63 \rightarrow 8'hE4 \rightarrow 8'h3D$". If the Trojan is triggered, the fifth output bit is forced to logic 1. Thus, the Trojan activation probability is $\frac{1}{2^{32}}$

The next class of Trojans inserted by the red team degrade the reliability of the IC. For example, a Trojan can increase the operating temperature of the IC. Since, increase in temperature accelerates aging [12], the Trojan-infected IC will degrade faster than a Trojan-free IC. The red team designed two such reliability degrading Trojans (Trojans 5 and 6).

(*v*) *Trojan 5* degrades the reliability of an IC. It consists of a trigger part and a payload part. The Trojan is triggered by a special input, $8'bxxx1xxxx$. Its activation probability is $\frac{1}{2}$. The payload consists of fifty 17-stage ring oscillators. Upon activation, this Trojan increases the power consumed by Design A by 20%. Consequently, the temperature of the IC is increased which accelerates aging.

(*vi*) *Trojan 6* is similar in functionality and the structure to Trojan 5. The special input is $8'b01011100$. Consequently, the activation probability is $1/256$. The payload consists of twenty 17-stage ring oscillators. Upon activation, it increases the power consumed by Design A by 9%.

The red team design Trojans 7 and 8 to evaluate the ability of detection techniques against Trojans whose sizes are less than 1% of the total design.

(*vii*) *Trojan 7* uses an inverter. It modifies the value in output pin and has 0.13% area overhead.

(*viii*) *Trojan 8* uses of 11 buffers and has 1.4% area overhead.

C. Competition

The red team designed six variants of Design A and two variants of Design B. Each variant is infected with a Trojan. For each variant (Trojan), the red team provided two Trojan-free bitstreams and two Trojan-infected bitstreams. To mimic the effects of process variations, the bitstream of a variant is modified by mapping the logic to a different set of LUTs.

The blue team is required to identify the Trojan-infected bitstreams. To mimic the real world scenario where a designer has access to the Trojan-free design, the Trojan-free netlists of Design A and Design B are provided to the blue teams. Seven teams from six different universities participated as blue teams.

III. Results and analysis of Trojan detection techniques

In this section, we will analyze the effectiveness of different Trojan detection techniques employed by the blue teams. The seven blue teams employed three different Trojan detection techniques: Trojan activation, power-based side-channel analysis, and delay-based side-channel analysis. The different methods employed different blue teams are listed in Table I.

TABLE I
Trojan detection techniques used by different blue teams

	Activation	Power	Delay	Others
Team 1	✓	✓	✓	✗
Team 2	✓	✓	✓	Electromagnetic
Team 3	✓	✓	✗	✗
Team 4	✓	✓	✓	✗
Team 5	✓	✓	✗	✗
Team 6	✓	✓	✗	✗
Team 7	✗	✓	✗	✗

A. Trojan activation

All the blue teams, except one, used Trojan activation method. Test patterns were generated, applied to the target bitstream and the golden design, and the corresponding outputs are compared against each other. If the outputs match, then that target bitstream is classified as Trojan-free. If the outputs do not match, then that target bitstream is classified as Trojan-infected.

Different blue teams used different techniques to generate the test patterns. For instance, one blue team used "pseudo-Hadamard" and "Fibonacci transforms" to generate pseudo-random test vectors. They successfully activated T1, T2 and T3 in Design A and T1 in Design B. Another blue team enumerated all the 256 different input vectors of Design A. They were able to detect Trojan 1 and Trojan 2 since these Trojans were triggered by one input vector. However, they were not able to detect Trojan 3 and Trojan 4 as they require sequence of input vectors to get triggered.

However, this approach suffers from several limitations. First, for a design with N inputs, there are 2^N possible inputs. Hence, applying exhaustive set of input patterns is impractical. Second, this method can trigger only combinatorial Trojans. In case of sequential Trojans, one has to apply all possible sequences of input patterns, which is impractical. Thus, this method cannot ensure triggering sequential Trojans. Lastly, this method can detect only Trojans that produce wrong outputs. It will not detect Trojans that degrade performance, reliability or leak secret information. Thus, as expected, no blue team was able to detect Trojans 5, 6, and 8 using this method.

B. Power-base side-channel analysis

Since Trojans consume additional power beyond that consumed by the original design, blue teams attempted to detect this additional power consumption. To measure the power, one blue team placed a "current-sensing" resistor between the power supply and the power port of the FPGA, and measured

TABLE II
HARDWARE TROJAN DETECTION RESULTS FROM DIFFERENT TEAMS. D (SUCCESSFULLY DETECTED), U (UNDETECTED).

	T1	T2	T3	T4	T5	T6	T7	T8
Team 1	✓	✓	✓	✓	✓	✓	✓	✓
Team 2	✓	✗	✓	✓	✓	✓	✓	✓
Team 3	✗	✓	✓	✓	✓	✓	✓	✗
Team 4	✓	✓	✓	✓	✓	✓	✗	✗
Team 5	✓	✓	✓	✓	✓	✓	✗	✗
Team 6	✓	✓	✓	✓	✓	✓	✓	✓
Team 7	✓	✓	✗	✓	✓	✓	✗	✓

the voltages drop across the resistor using an oscilloscope. Another team measured the current flowing through the external power supply.

Blue teams performed leakage and dynamic power side-channel analysis. Leakage/static power analysis was performed by applying an input vector and measuring the power consumption once the output gets settled down. Dynamic power analysis was performed by consecutively applying two input vectors and measuring the power consumption during the transition.

Leakage power analysis was able to detect Trojans that are usually big (>1% of the total design). For instance, Trojan 8, which has 11 buffers, was detected by this method. Dynamic power analysis was able to detect sequential Trojans (Trojan 3 and Trojan 4) and Trojans that degrade reliability (Trojan 5 and Trojan 6). Since the D-flip flops in the triggering part of the sequential Trojans switch during the transition of input patterns, these Trojans were exposed to dynamic power analysis. The reliability degrading Trojan used ring oscillators that switch at very high speed, thereby consuming a lot of dynamic power. Consequently, they were detected by dynamic power side-channel analysis.

C. Path Delay

Trojans increase the delays of circuit paths to which they are connected to. Such Trojans can be detected by measuring the path delay from a primary input or pseudo input (D flip-flop) to a primary output or pseudo output (D flip-flop). If the delay in the target IC is different from that in the golden design, then that IC is considered Trojan-infected. Path delay can be measured by causing a transition at the input, observing a transition at the output, and calculating the delay between the two transitions. Measuring path delays in combinational designs, such as Design A, is relatively easy when compared to measuring path delays in sequential designs. In the case of sequential designs, path delays are calculated by sensitizing the path and increasing the clock frequency until the flip-flops latch a wrong value. This clock frequency will indicate the delays of the sensitized paths. However, for short-delay paths, this clock frequency is too high to be realized on the FPGA. Thus, not many team used this method, as shown in Table I

In this competition, path delay-based methods were not as effective as activation- and power-based methods, as they required test vector generation to sensitize critical and non-critical paths, and equipment like oscilloscope to measure the delay. In addition, most teams generated random patterns to sensitize the paths. Consequently, they were not able to sensitize all the paths in a circuit. Furthermore, the red team inserted Trojans (except Trojan 8) in short-delay paths. Measuring such paths is impractical on an FPGA as they require clock frequencies that are higher than the maximum available clock frequency on the target platform (50MHz is the maximum frequency).

Path-delay based detection method was able to detect Trojans that had a large impact on path delay. For instance, Trojan 8, which has 11 buffers, was detected by this method. However, Trojan 7, which has only one inverter, was not detected by this method as it had negligible impact on path delay.

IV. CONCLUSIONS

In summary, activation-based Trojan detection technique is an effective approach to detect combinational Trojans that get triggered by input vectors and corrupt outputs. Unfortunately, they are not able to detect sequential Trojans and require a large volume of input patterns. Static and dynamic power side-channels are widely used by several teams and are successful in detecting Trojans. Furthermore, the blue teams are able to craft their input patterns to aid their side-channel analysis. Delay side-channel analysis is not attractive to many blue teams as it is difficult to implement. The main reason for this limitation is the relatively low clock frequency available on the target platform. The Trojans designed in the competition will be used a part of the dynamic trust benchmark circuits at TrustHub [13].

V. ACKNOWLEDGMENT

This work was supported in part by The National Science Foundation (0958510,1059328), Army (W911NF-11-1-0470), Air Force Research Labs and Intel.

REFERENCES

[1] "High Performance Microchip Supply," http://www.acq.osd.mil/dsb/ADA435563.pdf.
[2] "List of Accredited Suppliers," www.dmea.osd.mil/otherdocs/Accredited Suppliers.pdf.
[3] X. Wang, H. Salmani, M. Tehranipoor, and J. Plusquellic, "Hardware Trojan Detection and Isolation Using Current Integration and Localized Current Analysis," *Proceedings of the IEEE International Symposium on Defect and Fault Tolerance of VLSI Systems*, pp. 87–95, Oct. 2008.
[4] Y. Jin and Y. Makris, "Hardware Trojan detection using path delay fingerprint," *Proceedings of the IEEE International Workshop on Hardware-Oriented Security and Trust*, pp. 51–57, Jun. 2008.
[5] S. Narasimhan, D. Du, R. Chakraborty, S. Paul, F. Wolff, C. Papachristou, K. Roy, and S. Bhunia, "Multiple-parameter side-channel analysis: A non-invasive hardware trojan detection approach," pp. 13–18, 2010.
[6] H. Salmani, M. Tehranipoor, and J. Plusquellic, "A novel technique for improving hardware trojan detection and reducing trojan activation time," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 1, pp. 112–125, 2012.
[7] J. Rajendran, V. Jyothi, O. Sinanoglu, and R. Karri, "Design and analysis of ring oscillator based design-for-trust technique," *Proceedings of the IEEE VLSI Test Symposium*, pp. 105–110, May 2011.
[8] A. Ferraiuolo, X. Zhang, and M. Tehranipoor, "Experimental analysis of a ring oscillator network for hardware trojan detection in a 90nm asic," pp. 37–42, 2012.
[9] R. Karri and J. Rajendran, "Embedded systems challenge," isis.poly.edu/~jv/esc.
[10] R. Karri, J. Rajendran, K. Rosenfeld, and M. Tehranipoor, "Trustworthy Hardware: Identifying and Classifying Hardware Trojans," *IEEE Computer Magazine*, vol. 43, no. 10, pp. 39–46, oct. 2010.
[11] http://www.digilentinc.com/Products/Detail.cfm?Prod=BASYS2.
[12] R. Cowie, John M. G.; Ferguson, "Physical aging studies in poly(vinylmethyl ether). i. enthalpy relaxation as a function of aging temperature," *Macromolecules*, vol. 22, pp. 2307–2312, 1989.
[13] TrustHub, "http://trust-hub.org/."