The Undetectable and Unprovable Hardware Trojan Horse

Sheng Wei Miodrag Potkonjak
Computer Science Department
University of California, Los Angeles (UCLA)
Los Angeles, CA 90095
{shengwei, miodrag}@cs.ucla.edu

ABSTRACT

We have developed an approach for automatic embedding of customizable hardware Trojan horses (HTHs) into an arbitrary finite state machine. The HTH can be used to facilitate a variety of security attacks and does not require any additional gates, because it is morphed into the specified design. Even after the HTH induces provable damage, one is not capable of proving that any malicious circuitry is embedded into the design. The main ramification of the developed HTH is that hardware and system techniques should move from HTH detection toward synthesis for trusted systems.

1. INTRODUCTION

Hardware Trojans horses (HTHs) [1] are malicious modifications to integrated circuits (ICs) during design (e.g., by untrusted CAD tools) or manufacture (e.g., by untrusted foundries). Recently, HTH has posed great concerns with regard to the security and integrity of ICs with the rapid growth of IC outsourcing. The detection of HTHs has triggered a great deal of attention in the IC community, due to the fact that the consequences of HTHs can be extremely severe for security-sensitive systems.

The existing HTH research [1][2][3][4] targeted only on Trojans that are physically present and thus observable on the target IC, either in the form of additional malicious components or modifications toward the target circuit. Although these types of HTHs can be well hidden under the target circuit, the difficulty level for detection is limited due to the following two reasons. (1) the embedded HTH would result in at least one type of variation in the observable properties of the IC, including but not limited to physical structures (e.g., layout and wiring) and side channels (e.g., delay and power); and (2) the HTHs under consideration are identical on different chips of the same design due to the high cost of customizing the design and manufacturing for HTH insertion. As a result, once one chip compromised by HTHs is detected during the IC test, all other chips under attack can be easily identified in a straightforward way.

DAC'13, May 29 - June 07 2013, Austin, TX, USA.

We argue that it is completely feasible to create challenging HTHs and bypass the existing detection schemes that are subject to the above limitations. It is rather important to investigate on these HTH attacks and motivate security primitives from a completely new angle. Based on these thoughts, we develop a zero-overhead, customizable HTH model that an attacker could leverage to create untrusted CAD tools and trigger undetectable security attacks. Our undetectable HTHs have the following features:

Zero-overhead. Our HTH model leverages the redundant states (called HTH states) in the finite state machine (FSM) of the target circuit for security attacks, which does not require any additional hardware to trigger the HTH during normal IC operations and, therefore, exposes no observable variations in the IC properties.

Customizable. The proposed HTH model induces different and customizable security attacks on different ICs without introducing additional manufacturing costs. Consequently, even if one instance of the HTH is detected, it is extremely difficult for the detection procedures to prove the presence of HTHs, nor can they generalize the found instance to other chips under test. We achieve this goal by employing post-silicon device aging caused by the negative bias temperature instability (NBTI) effect [5]. The attacker could intentionally age the target ICs after manufacture in such a way that unpredictable delay faults occur at runtime to transition the IC from normal states to the HTH states.

Therefore, the main ramification of our proposed HTH model is that it forces the detection mechanisms to move from traditional detection to sequential synthesis. Not only the cost for detection is significantly increased, but also the fundamental paradigms in the existing detection approaches have to be revisited and reconsidered in order to achieve reliable HTH detection schemes.

2. MOTIVATIONAL EXAMPLE

Figure 1 shows a motivational example of our proposed undetectable HTH model. Figure 1(a) is the finite state machine of a mod-3 up/down counter, including two inputs (x_1, x_0) that control the counter to stop, count up, and count down; and 4 states that can be implemented by 2 flip-flops. Among all 4 states, only 3 of them are valid states of the counter, i.e., representing the count number 0, 1, and 2. The shaded state S_3 is a redundant state (or don't-care state) that cannot be reached from any other states using any inputs. An attacker could leverage S_3 to trigger a variety of attacks, such as leaking confidential information or consuming higher energy.

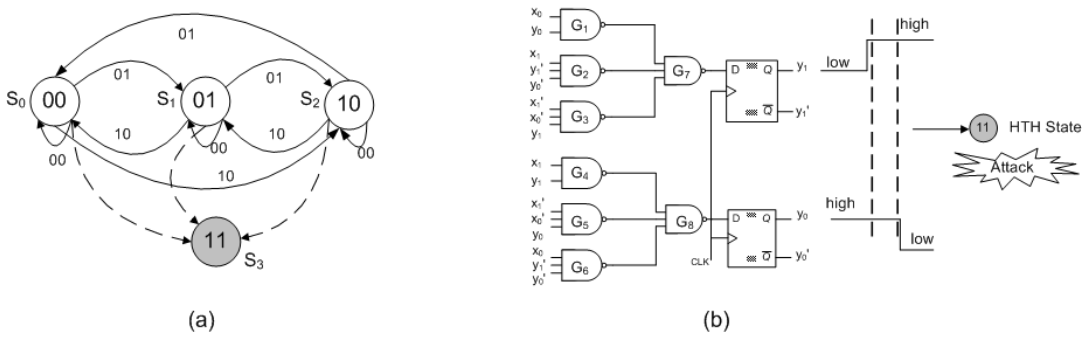

Figure 1: Motivational example of the undetectable hardware Trojan horses: (a) finite state machine of a mod-3 up/down counter, which includes 3 normal states (i.e., S_0, S_1, and S_2) and 1 redundant state (i.e., S_3); and (b) demonstration of the HTH state transition using device aging.

Figure 1(b) demonstrates the design of the sequential circuit based on the FSM, which shows the transition from normal states (i.e., states S_0, S_1, and S_2) to the don't-care state (i.e., state S_3), so that the desired HTH attack can be triggered at runtime. Our approach is to intentionally age (i.e., stress the corresponding transistors) a certain set of gates and trigger delay faults at the circuit output. For example, the attacker could age gate G_8 so that the signal transmitted to F_2 is delayed. It is possible that the delayed signal for F_2 causes delay fault, e.g., both F_1 and F_2 stay at signal 1, which transitions the circuit into the HTH state.

The trigger of delay fault and thus the state transition is fully customizable, in the sense that the attacker can selectively age different components for different chips post-silicon, which transitions the target circuit to different HTH states from different normal states. Even in small designs, there are exponentially many combinations of transitions that can be leveraged by the attacker to complicate and obfuscate the attacks.

3. FEASIBILITY STUDY AND VALIDATION

The feasibility of the proposed HTH attack is based on the assumption that there are large numbers of redundant states available in the target circuit. We argue that the assumption holds for the following two reasons. Firstly, the design of modern sequential ICs often results in large numbers of redundant states for the consideration of performance and ease of integration. Secondly, even if the original design specification does not indicate enough don't-care states, the attacker could easily minimize the FSM [6] to create equivalent designs that include many redundant states.

4. HTH DETECTION REVISITED

As a consequence of the undetectable HTH model, the traditional HTH detection mechanisms have to be revisited to accommodate the elevated difficulty level for ensuring a trusted IC system. In order to achieve this goal, we argue that the current HTH detection approaches [1], which rely on the monitoring of the end system in the post-silicon stage, have to be moved to sequential synthesis at the design time. In other words, the detection process must examine the redundant states generated by the untrusted CAD tools and exclude the possibility of HTH attacks early at the design time, which is an extremely difficult task.

Our idea to address the problem is to enforce a specified system at design time, where all or a part of the don't-care states are either explicitly removed or incorporated as a well defined state. In this way, we can limit the freedom of manipulating the FSM that is exposed to the untrusted tools. The downside of this solution is that it may compromise the performance gains obtained from the don't-care states. Therefore, a careful design is required to balance the tradeoff between performance and security of the system.

5. CONCLUSION

We have developed a zero-overhead and customizable hardware Trojan horse that cannot be detected by existing HTH detection approaches. The attack model leverages redundant states in the finite state machine and triggers the malicious state transition using device aging. We show that the HTH detection schemes must move from post-silicon monitoring to complex design-time synthesis in order to capture the new attack. By presenting the new HTH model, we aim to motivate new HTH detection research in the community to ensure the security and integrity of the ICs.

6. ACKNOWLEDGEMENTS

This work was supported in part by the NSF under Award CNS-0958369, Award CNS-1059435, and Award CCF-0926127, and in part by the Air Force Award FA8750-12-2-0014.

7. REFERENCES

[1] M. Tehranipoor, F. Koushanfar, A Survey of Hardware Trojan Taxonomy and Detection, IEEE Design and Test of Computers, Vol. 27, No. 1, 2010, pp. 10-25.

[2] S. Wei, S. Meguerdichian, M. Potkonjak, Gate-level Characterization: Foundations and Hardware Security Applications, DAC 2010, pp. 222-227.

[3] S. Wei, K. Li, F. Koushanfar, M. Potkonjak, Hardware Trojan Horse Benchmark via Optimal Creation and Placement of Malicious Circuitry, DAC 2012, pp. 90-95.

[4] S. Wei, L. Kai, F. Koushanfar, M. Potkonjak, Provably Complete Hardware Trojan Detection Using Test Point Insertion, ICCAD 2012, pp. 569-576.

[5] H. Baba, S. Mitra, Testing for Transistor Aging, VTS 2009, pp. 215-220.

[6] L. Yuan, G. Qu, Information Hiding in Finite State Machine, IH 2004, pp. 340-354.