A Survey of Hardware Trojan Detection, Diagnosis and Prevention

He Li †, Qiang Liu †, Jiliang Zhang ‡*, Yongqiang Lyu §
†School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China
‡Software College, Northeastern University, Shenyang 110819 China
§Research Institute of Information Technology, Tsinghua University, Beijing 100084 China
{heli; qiangliu}@tju.edu.cn, zhangjl@swc.neu.edu.cn, luyq@tsinghua.edu.cn

Abstract—Hardware Trojans (HTs) can be implanted in security-weak parts of a chip with various means to steal the internal sensitive data or modify original functionality, which may lead to huge economic losses and great harm to society. Therefore, it is very important to perform hardware Trojan detection and diagnosis, finding potential safety hazards and apply protection techniques in the whole IC design cycle, in order to enhance the security of chips. In this paper, we elaborate an IC market model, and describe the potential HT threats faced by the parties involved in the model. Then we survey the recent research advances in the countermeasures against HT attacks, which are classified into HT detection, diagnosis and prevention. Finally, the challenges and prospects for HT defense are illuminated.

Index Terms—Hardware Trojan detection; hardware Trojan diagnosis; hardware Trojan prevention; IC market model

I. INTRODUCTION

In recent years, a new type of hardware attack for integrated circuits (ICs), called hardware Trojan (HT), has become a major concern in the IC industry. ICs become untrustworthy when HTs are inserted into them by adding a piece of circuit or modifying the designs for malicious purposes. HTs can make serious attacks such as disabling or altering the functionality of an IC, or leaking sensitive information like cryptographic keys embedded in the chip. Therefore, it is urgent to conduct research on HT detection, diagnosis, and prevention.

Several excellent surveys on hardware Trojan attacks and countermeasures have been published recently. To begin with, we provide a short description of these works and define the scope of the present survey. In the first survey [1] on malicious circuits, Wang et al. elaborated both HT taxonomy and Trojan detection methods. Tehranipoor et al. [2] presented a more in-depth discussion and classification of HT attacks, covering three topics: Trojan design and taxonomy, Trojan detection methods and design for hardware trust. Later, a more comprehensive review [3] augmented the complex HT threat models and illustrated the feasible countermeasures in specific fields concerning HT attacks. Another survey [4] discussed the feasibility of Trojan insertion at each stage of IC development and production chain, and presented how different stages in the development process give an adversary opportunities to maliciously modify ICs.

Compared with the existing surveys, this survey from a different angle uses an IC market model to elaborate specific HT threats faced by the parties involved in the model. The present paper reports a new progress in HT detection approaches, especially in HT diagnosis which not just detects the existence of HTs but also finds their locations. In addition, the up-to-date HT prevention and real-time monitoring approaches are also surveyed in this paper. The purpose of this paper is to meet the demand for a survey on the state of HT attack countermeasures with an emphasis on the recent developments that have taken place within the past three years and a focus on the approaches that most likely will be reduced to practice.

The remainder of this paper is organized as follows. In Section II, we present the preliminaries about hardware Trojan and describe the IC market model, as well as HT threats faced by the parties involved in the model. In Section III, we survey the countermeasures for HT attacks, including HT detection, diagnosis and prevention approaches. Challenges and prospects are presented in Section IV. Section V concludes the paper.

II. HARDWARE TROJAN THREAT AND VULNERABILITIES

A. Preliminaries about Hardware Trojan

Hardware Trojan is a piece of circuit that is added to the design or modified from the original design for malicious purposes. Figure 1 shows a typical model of HT, which contains trigger, HT circuit, and payload [5, 6]. To be hidden in chips, the HTs usually are designed to be silent in most of time. The trigger is associated with rare signals or events [3]. When the specified signal or event appears, the HT circuit is activated to be functional. The payload circuit is responsible for implementing HT attacks, which may result in serious effects such as denial of service (DoS), confidential information leakage, and chip reliability degradation. Various HTs have been designed with different activation mechanisms. A comprehensive survey on the taxonomy of HTs can be found in [3].

B. IC market model

In this section, we will introduce an IC market model in a way similar to the FPGA-based system market model [7], as shown in Figure 2. Typically, there are five parties involved in the IC design, manufacture, and application flow. The role of each party is described below.
C. HT threats in IC market model

This section analyzes the existence of HT threats during the interactions between every two parties involved in the model.

1) HT threat between SoC designers and foundries: during the fabrication process, there is no guarantee that foundries do not insert a certain type of HT in the chips. Chips fabricated in foundries may be threatened by untrusted staff or third parties to whom the fabrication process is accessible. For example, a Trojan can be implanted into the IC by intentionally/unintentionally modifying the dopant level or the mask layout either during the sample or mass production \[4\]. In addition, foundries have their confidential instruments to manipulate the chip fabrication for some malicious purposes and may outsource the mask generation to a third party which has the opportunity to maliciously include mask macros in the GDSII.

2) HT threat between SoC designers and IP vendors: in the interaction between SoC designers and IP vendors, the former needs to ensure that the acquired IPs do not hide malicious function units which will be extremely difficult to be detected afterwards. Several different types of Trojan can be designed by a proficient adversary during the pre-silicon stage. An untrusted insider in IP vendors can easily manipulate the RTL and insert malicious codes, modify macros during the design synthesis, and even alter the placement&route so as to make room for the Trojan circuitry \[4\]. Furthermore, an untrusted contractor who develops parts of the specification for IP vendors and SoC designers can include malicious elements.

3) HT threat between IP vendors (or SoC designers) and EDA vendors: EDA tools are widely used in many critical design stages. The software tools developed by EDA vendors may contain malicious codes which will collect valuable data in IPs/SoCs. Recently, Qu and Yuan \[8\] analyzed the security vulnerabilities in EDA design tools and reported that logic implementations deduced by EDA tools may do more than required, regardless of the trustworthiness of the design team, the source of the EDA tools, and IP providers. This unexpected breaches could be exploited by adversaries to fulfill attacks.

4) HT threat between End users and SoC designers: end users who do not have a chip-level design team are concerned with HT attacks in products purchased from SoC designers. The HTs can be embedded into the chips during the SoC design step, to bypass the software security facilities and spy the users. The end users practically have no ability to detect this kind of hardware-level security threat, making them do not trust the traditionally reliable chips any more. The evidence of HTs/backdoors hidden in weapons control systems, nuclear power plants, and public transportation systems has been reported in \[9\].
III. COUNTERMEASURES FOR HARDWARE TROJAN THREAT

Based on HT threats discussed in Section II, we summarize current countermeasures against the threats existing in every two parties in the IC market model in Table I and then presented the countermeasure approaches in the following.

In order to find the security threats in the IC market model and prevent successful HT attacks, various countermeasures have been developed. In this paper, countermeasures against HT attacks are classified into three categories: 1) HT detection approaches, 2) HT diagnosis approaches and 3) HT prevention approaches. HT detection is the process that determines whether any HTs exist in the circuit. HT diagnosis is to determine the location of HTs in the circuit, so that one can either remove or mask the HTs from the circuit. HT prevention approaches mostly apply to the design stage to increase the difficulties of HT insertion or to prevent the successful HT insertion in IC development.

A. Approaches for hardware Trojan detection

1) HT detection in pre-silicon stage: can be used to detect the HTs inserted by the EDA tools or brought in the IP cores. Adversaries always try to insert HTs in a way such that the HTs are dormant during functional verification. Therefore, the HTs are resistant to the traditional functional verification approaches. Recently, several trust verification approaches have been proposed to flag suspicious circuits inserted in the design stage. These techniques exploit formal verification and functional simulation methods.

The first set of approaches uses static functional verification techniques such as formal verification and assertion-based verification. The design comparison method [10] is proposed to resolve a question “How does one verify that a block does what it is expected to do, and nothing else?” The basic idea is to make a comparison between two blocks from different resources with equivalent functionality. The full process involves wrapping designs and unrolling internal states to express each output entirely in terms of past and present inputs, then completely removing state components such as flip-flops and latches, leaving only combinational logic and delayed inputs, and finally comparing the designs with a Boolean satisfiability (SAT) solver to find redundant logic.

Zhang et al. [11] define all functions in the specification as properties, and the corresponding assertions are defined simultaneously. Then, coverage metrics (code coverage and functional coverage) are analyzed to identify uncovered parts, which are considered as suspicious circuits. Moreover, a verification approach of system specification and implementation is also presented to identify extra functionality in hardware designs [12].

Functional Analysis for Nearly-unused Circuit Identification (FANCI) is proposed in [13] to identify input signals with weak effect by static Boolean function analysis. The criteria for suspicious inputs is define as the control value shown in Eq.(1).

Control Value \( w_1, w_2 \) = \( \frac{\text{counter}(w_1)}{\text{size}(w_2)} \) (1)

where \( \text{counter}(w_1) \) denotes the total number of rows of \( w_1 \) which determines the value of output \( w_2 \) in the truth table; \( \text{size}(w_2) \) denotes the total number of rows of \( w_2 \) in the truth table. Take a multiplexer shown in Figure 3 as an example. Input \( A \) has a control value 0.25, which can be obtained by counting the number of rows with same value in Column \( A \) and \( M \) and the total number of rows of Column \( M \). For practical calculation, we only need to look through a half of the truth table, because the two halves represent the same property. For a malicious multiplexer, there are 64 additional select bits. When those 64 bits match a specific 64-bit key, \( M \) is changed to a malicious payload. However, each additional bit only affects a small fraction of Column \( M \) and their control value is \( 2^{-65} \). After calculating the control value for all inputs, the approach in [13] derives a threshold for control value and those inputs with the value below the threshold are considered as suspicious inputs.

The second set of approaches combines static and dynamic verification techniques. Hicks et al. [14] first formulate the HT detection as unused circuit identification (UCI) problem which can be considered as suspicious circuits, whenever they do not affect outputs during simulation. The UCI algorithm traces all signal pairs, and selects those signals with equal properties as suspicious HT insertion targets. Afterwards, the suspicious
circuit is isolated and a exception notification logic is added to notify the abnormalities at runtime.

Later, an optimized approach VeriTrust [15] flags suspicious circuits by identifying potential trigger inputs used in parasite-based HTs. A parasite-based HT exists along with the original circuit, and does not cause the original design to lose any normal functionalities. For example, the K-map analyzed in [15] is shown in Figure 4. The K-Map of the parasite-based HT-inserted circuit is shown in Figure 4(b), where the third row represents the malicious function while other rows show the normal function. By comparing it with the K-Map of the original circuit in Figure 4(a), the parasite-based HT enlarges the K-Map size with additional inputs so that it can keep the original function while embedding the malicious function. If we set all entries of the malicious function as don’t cares, the trigger inputs (i.e., \( t_1 \) and \( t_2 \)) become redundant. These redundant inputs are then flagged as potential HT trigger inputs for the further examination [16].

With the system complexity increasing, computational effort for verification methods rises dramatically [17]. It is not proper to carry out formal verification to the whole circuit. A metric for the assessment of Trojan inexistence called Trojan Assurance Levels (TAL) [18] is introduced to locate the insecure area of the chip design. This metric can be mathematically defined by evaluating the circuit functionality, structure and functional interactions at different levels of abstraction. It is expected that the HT detection process will be more efficient by focusing on the regions with high possibility of HT inserted according to TAL.

2) HT detection in post-silicon stage: can be used in the post-silicon testing process to find the HTs inserted during the design stage and the manufacturing stage. The detection approaches can be further divided into side channel analysis and logic test.

a) Side channel analysis: side channel (SC) analysis has been widely applied to HT detection due to the fact that the inserted HTs would have effects on the circuit’s power consumption [19–25] and signal delay [26–29].

According to current analysis, HT embedded in the original circuit will add extra leakage current and power, but it is difficult to be detected directly due to the tiny impact on the whole circuit. Therefore, various design partition-based approaches have been proposed [22, 23, 25]. The basic idea is to augment the effect of the HTs. For instance, a scan cell distribution based partition technique [25] is proposed to divide the circuit into regions. Then, activity-driven test pattern is generated to magnify the activity in the target region where the HT may be located. Finally, power ports are placed in each region to measure the localized transient current anomalies for HT detection.

Compare with other side-channel signals, path delay has its advantage: each path delay is independent with each other and they can be measured separately. However, the main challenge is that the tiny impact on HT delay cannot be effectively measured under the increasing level of process variation. Hence, Cha and Gupta [27] focus on reducing the process variation influence on Trojan delay by calibrating the effect of process variation and building the delay model for each logic blocks, and then carry out HT detection in each block. In their further research, the additional path delay induced by Trojan is maximized by using a path selection scheme [28].

b) Logic test: as semiconductor technique advances, side channel analysis-based detection approaches alone become ineffective due to the significant impacts of process variation. It is suggested to combine side channel analysis with logic test approaches, which focus on generating appropriate test patterns for HT detection [3].

Since an adversary can insert a plurality of HT instances, it is impractical to enumerate all possible Trojan instances to generate deterministic test patterns and measure Trojan coverage. Statistical approach for test vector generation have been developed to address this issue.

A random sampling approach, MERO (Multiple Excitation of Rare Occurrence) [30] is proposed to generate effective input vectors. The basic concept is to detect low probability conditions at the internal nodes and then derive an optimal set of vectors to activate the rare nodes at least \( N \) times, in a similar way to N-detect test used in stuck-at ATPG. The probability of activating a Trojan is improved by increasing the transitions of nodes that are random-pattern resistant. However, the process of test vector generation is time-consuming.

To improve the efficiency of HT detection, Li and Liu [31] first design a HT detection acceleration approach based on signal word-level statistical properties with mean (\( \mu \)), standard deviation (\( \sigma \)) and autocorrelation (\( \rho \)). Compared with existing HT detection methods, this acceleration technique can dramatically enhance the rare nodes transition activity in order to increase the probability of activating HTs, and less detection time will be required. The enhancement of rare nodes transition in a 5-tap FIR circuit is presented in Figure 5. Transition activity of the \( 6^{th} \) bit can be increased by statistical signals with different (\( \mu, \sigma, \rho \)).
Fig. 4. K-map in VeriTrust [15]. (a) K-map of original circuit; (b) K-Map of parasite-based HT; (c) K-Map of parasite-based HT in (b) by setting entities of the malicious function as don’t cares.

Fig. 5. Acceleration of Hardware trojan detection based on word-Level statistical properties management [31].

B. Approach for hardware Trojan diagnosis

As mentioned previously, various HT detection approaches have been developed to determine whether HTs exist in a circuit. However, for the IP vendor or SoC designers, they prefer to acquire the specific HT information in their product in terms of HT types, HT locations and HT triggers, in order to remove the threats.

Wei and Potkonjak proposed approaches based on circuits segmentation methods for HT diagnosis. Their scalable HT diagnosis approach based on circuit segmentation and gate level characterization (GLC) is proposed in [22]. The basic procedure, as shown in Figure 6, contains three phases: segmentation, HT detection and diagnosis, and post-processing. A segmentation model is first trained by segment properties including controllability ratio, correlation ratio and GLC accuracy. The model is then proposed to divide large circuits into small sub-circuits. Afterwards, in sub-circuits with small number of gates it is easier and more accurate to detect and diagnose HTs by tracing leakage power. In the third phase, statistical methods are applied to validate the prediction results in the post-process. The whole process is repeated multiple times if necessary.

Another HT diagnosis approach proposed in [23] is based on segmentation and consistency analysis of gate-level properties. One can firstly detect the HTs by measuring the gate-level properties in two segments with overlapping gates. These gates can exhibit inconsistent leakage power. Then a third segment with the same overlapping gates is introduced to indicate the HT’s position. Figure 7 shows an example of the consistency-based HT diagnosis. There are three segments in the circuit. Segment 1, Segment 2 and Segment 3 with the same overlapping gate \( X \) have leakage power scaling factor \( S_1 \), \( S_2 \), and \( S_3 \), respectively. Given that \( S_3 \) has the same value with \( S_1 \), it indicates that Segment 2 potentially contains a HT.

C. Approach for hardware Trojan prevention

The HT detection and diagnosis approaches, though promising, still face some challenges such as rare node identification, process variations, and measurement deviation. To improve the effectiveness of these approaches, ICs must be designed with self-protection awareness. Currently, obfuscation, layout-filler, dummy circuit insertion and split manufacturing are the main techniques for HT prevention.
In the future, we think it is necessary to build a trusted hardware-oriented security and trust system. Most existing HT detection techniques rely on the Golden HT diagnosis approaches are a prospective research field, and prevention in the whole IC market model, it is still necessary to construct a last defender to realize on-chip monitoring during runtime [3, 37]. Once the abnormal operation of the circuit happens, alert mechanism will shut the circuit function, and trigger other security measures to prevent further consequence caused by hardware Trojans. Analog sensors such as thermal sensor can also be exploited to detect deviations in power/thermal profiles caused by Trojan activation [37]. Runtime monitor has been comprehensively introduced in [3], where runtime monitor techniques are classified into three sub-classes: configurable security monitors, variant-based parallel execution, and hardware-software approach. It is believed that the detection at the chip testing phase and run-time monitoring are complementary to detect Trojans.

**IV. Challenges and Prospects**

With advanced technologies, adversaries are likely to inflict new and unanticipated attacks which are difficult to be tackled by existing countermeasure approaches. Therefore, countermeasure techniques against HT attacks need further development, in order to win the race. The possible challenges and prospects in this field are listed as follows.

- Systematic and time-efficient HT detection approaches for third-party IPs at pre-silicon stage are demanded.
- Vulnerabilities from EDA tools are not widely concerned currently. Therefore, there is an underlying risk for the whole IC design cycle.
- With the increasing size of ICs, an adversary can exploit a large number of Trojan instances in various forms and sizes [38]. It can be extremely challenging to activate arbitrary Trojan instances and observe their effects in advanced technologies with process variations [3].
- Most existing HT detection techniques rely on the Golden chip which is difficult to obtain, even does not exist. Therefore, HT detection and diagnosis without reference model could be in an urgent demand.
- HT diagnosis approaches are a prospective research field, but accurate orientation of HTs is very difficult for large and complicated circuit designs.
- It is suggested that combining HT detection, diagnosis, prevention and runtime monitoring will probably provide a complete solution to address the HT issues [3].
- In the future, we think it is necessary to build a trusted third party (TTP) with all necessary equipments and techniques to focus on the HT detection for end users.

V. Conclusion

In this survey, we elaborate an IC market model, and describe the HT threats at the interactions between parties involved in the model. We survey HT detection, diagnosis and prevention approaches against the potential HT attacks. Finally, we discuss the challenges and the prospects for HT defense. In a word, tackling the hardware Trojan will require long-term and tough endeavor. With proper approaches, we could gradually increase the difficulty and cost of HT attacks and even eliminate them, and leave HTs to the past.

**Acknowledgment**

This work was supported by the National Natural Science Foundation of China under Grant No. 61204022, the Natural Science Foundation of Tianjin under Grant No. 12JCYBJC30700, Asia Research Center in Tsinghua University 2012 Young Scholar Program and the Central Universities Fundamental Research funding of China under Grant No. N120317003. The authors would like to thank Dr. Gang Qu for reviewing this article and providing us feedback.

**References**


[28] B. Cha and S. K. Gupta, “Trojan detection via delay measurements: A new approach to select paths and vectors to maximize effectiveness and minimize cost,”


