IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 9, 2020, accepted November 19, 2020, date of publication November 25, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040395

Dynamic Game for Strategy Selection in Hardware Trojan Attack and Defense

DAMING YANG, CHENG GAO, AND JIAOYING HUANG
School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
Corresponding author: Cheng Gao (gaocheng442@126.com)

ABSTRACT The offshore outsourcing introduces serious threats to semiconductor suppliers and integrated circuit(IC) users for the possibility of hardware trojans (HTs). To alleviate this threat, IC designers use the active defenses against HTs which are implanted by the malicious manufacturer. In this paper, a game-theoretic framework based on fuzzy theory is proposed to obtain the optimal strategy. It analyzes the interactions between the active defense designer and the malicious manufacturer. The attack and defense on IC is formulated as a noncooperative dynamic game. The active defense strategy is decided in IC design. And the overall payoff including design costs and losses avoided is optimized. Subsequently, the HT is implanted considering the implantation cost as well as the damage caused by it. To solve the problem of uncertain payoff caused by insufficient information, fuzzy variable is used to represent the influence of the defense coverage rate on the payoff. In order to verify the applicability of fuzzy variable in dynamic game framework, the existence of pure strategy Nash equilibrium in the game is proved. A solution algorithm for pure strategy Nash Equilibrium is proposed to obtain the optimal strategy of attacker and defender. Thanks to the case study of Field Programmable Gate Array (FPGA), the proposed framework is feasible for HT attack and defense game.

INDEX TERMS Hardware trojan, chip security, dynamic game.

I. INTRODUCTION

The tremendous advancements in semiconductor technology have resulted in the continuous compression of the time from integrated circuit (IC) design to market. This makes it impractical for an IC manufacturer to complete all the processes from design to manufacturing [1]. In addition, the trend of globalization is irresistible, leading to a novel industry model in which design, manufacturing, testing and packaging are separated and completed independently [2]. For the IC design company, it can't afford the high construction and maintenance cost of the production line with advanced manufacturing and processing capacity. It is an inevitable choice for the general design company to outsource the manufacturing to the third-party factory. With the popularity of this phenomenon, the threat to the IC security caused by the untrusted third-party foundry has attracted the attention of academic and industrial communities. An untrusted third-party foundry can implant malicious circuits into the original design without the designer's knowledge. This kind of malicious circuits is called hardware trojan (HT) [3].

HT is defined as a malicious inclusion/deletion/alteration, or an inadvertent design loophole of ICs or intellectual property (IP) cores. It can be used by the adversaries to achieve the malicious purpose as an attacker, which may cause disastrous consequences [2]. A HT can be implanted into an IC at manufacturing. It is inactive until activated by certain preset conditions. Once activated, it can attack the IC and cause serious consequences to the system. According to the attack effect, HTs can be divided into four types, which are functionality change, performance degradation, information leakage and denial of service [4]. To alleviate the potential security threats posed by HTs, substantial research efforts have been devoted to the studies of HT detections and design-for-security [5].

For the latency and concealment of HT, HT detections are limited by detection cost and process variation noise.Therefore, it is necessary to take targeted design-for-security in the IC design to reduce the threat of HT.It is defined as the approaches that make insertion of hard-to-detect HTs difficult or facilitate the detection during post-silicon validation [6]. The main way to solve this problem is to prevent HT from being implanted in the IC which is also known as the active deisgn [7]. Table 1 compares the properties of active defense with those of HT detection. Unlike HT detection as a pre/post-silicon method, active defenses is an IC design method as a pre-silicon method. While HT detection is used to determine whether the IC is HT

The associate editor coordinating the review of this manuscript and approving it for publication was Junggab Son.

TABLE 1. Differences between the HT detection and active defense.

HT Countermeasures	Application	Tools	Operating Principle	IC Resources Utilized
HT Detection	Pre/post-silicon	Testing Instrument	Verifying whether the designed or manufactured IC is implanted with HT	No
Active Defense	IC Design	EDA Tools	Preventing implantation or activation of HT	Yes

contaminated by testing instruments, active defense mainly prevents the HT implantation using EDA tools. Besides, different from the detection, active defense occupies the IC resources and effectively prevents implantation or activation of HT. Although the active defense is effective, it cannot completely prevent the HT from being implanted. The existing types of HT can be resisted by active defenses, but no one active defense can prevent all HT types [7]. This motivates the need for a mathematical framework to study the strategic interactions that occur between the HT designer and active defense designer. It anticipates the outcome of such interaction and provides the best strategy for defense designer.

Game theory studies the strategies and its equilibrium when the actions of decision-makers interact with each other. It is a mathematical method to study how the participants in competition acts to strive for the maximum benefits [8]. Meanwhile, it is good at describing the strategy dependence, which can well describe the attack-defense confrontation and the mutual influence between the attacker and the defender [9]. It provides a quantitative decision-making framework for the participants with conflicting interests, which is difficult to achieve by traditional security technology [10]. The remarkable success of game theory in a variety of research domains has inspired academic communities to explore its potential to address HT detections. Researches have focused on modeling the strategic interaction between an attacker and a detector in HT insertion/detection using game theory[27-31].

Different from the detection method, active defense occupies the IC resources as shown in Tab.1. This provides a payoff measure of the attacker and defender's actions. It is more practical than the aforementioned method using illustrative numbers to represent the action's payoff. For HT defense and attack, the active defense designer, as a defender, must decide on which defense to use in IC design and which trojans to resist, knowing all possible HT types that a malicious manufacturer can introduce. Subsequently, the malicious manufacturer, as an attacker strategically decides on which HT type to insert in IC manufacture while knowing the active defense. This stimulates the need for a dynamic game framework that allows for a better understanding of these strategic interactions between the attacker and the defender instead of the existing static game framework.

To meet this need, we propose a dynamic game framework for the HT implantation and active defense. The main contributions of this paper are summarized as follows.

1) ***Dynamic game for HT implantation and active defense.*** We formulate the problem as a noncooperative dynamic game which have two stages. In the first stage, the defender designs an active defense measure to resist HTs. The attacker inserts the HT in the second stage, and the action of the defender can be observed by the attacker before he acts. Both sides choose their own strategies to achieve the best payoff. This framework will provide the Pure Strategy Nash Equilibrium of the attack and defense in IC design and manufacturing.
2) ***Payoff modeling based on hardware resources.*** To be more practical, we use the cost-benefit function instead of illustrative numbers in related works to express the strategy payoff. We build a payoff model for HT defense decision-making based on Lee's model on network security, considering the effect of defense coverage on its payoff. The payoff is modeled based on hardware resources, design difficulty, attack effect and defense coverage.
3) ***Dynamic game based on fuzzy theory.*** The defense coverage rate is a quantitative index for HT defense and its impact on defense payoff is uncertain for insufficient information. To solve this problem, we use triangular fuzzy numbers to characterize the effect of defense coverage rate on defense payoff. On this basis, we prove the existence of pure strategy Nash equilibrium in fuzzy game and propose a method to solve the equilibrium.

This paper is organized as follows. Section 2 presents the background on past researches on game-theoretic approaches in HT detection and the motivation behind this work. The dynamic game framework based on fuzzy theory for HT attack and defense is described in Section 3. Section 4 presents the simulation results of attack-defense game on Field Programmable Gate Array (FPGA). Section 5 concludes the paper.

II. RELATED WORK

A. HT DESIGN

Since the HT was proposed in 2007, the related research in this field has never stopped [11]. The related technology at this stage has become a hot research topic. After more than ten years' efforts, both the attack and the defense technology of HT have been greatly developed. There are many kinds of HTs, and different types have different implementation methods.

In 2013, researchers from four universities in the United States uploaded more than 140 kinds of HTs to the trust-hub website, and built a preliminary HT library [12]. The host circuit of the samples in the library involves AES, RS232 and Ethernet, etc. They have been widely used by related researchers. HT designers can refer to their implementation methods, and the detection personnel can verify the effectiveness of the detection method by detecting these HTs.

Yang *et al.* [13] proposed a HT triggered by analog circuit in 2016. This trigger is extremely low-cost and secret, which only needs one gate at least. Before that, the researches on HT trigger design mostly focus on the digital circuit. He proposed an HT trigger with analog capacitor. Since then, people pay more and more attention to the HT with analog circuit.

In 2020, Subramani *et al.* [14] designed a HT to steal information with amplitude-modulating analog. He combined the digital circuit with analog circuit to realizes the leakage of the key information in the wireless encryption chip. It completed the long-range sustainable attack through amplitude-modulating technology.

With the researches of HT design, the HT library on Trust-hub website is constantly updated. According to the attack pattern, these HTs are divided into four types, which are Functionality Change, Information Leakage, Performance Degradation and Denial of Service. The case study of the proposed framework is against the HTs in the library.

B. ACTIVE DEFENSE FOR HT

Roy et al proposed a Hardware obfuscation method to lock the original design by randomly inserting additional gates and only a correct key makes the design to produce correct outputs [15]. To achieve the purpose of hiding the original design, MUX [16], gates [17], look-up table(LUT) [18], or Physically Unclonable Function(PUF) [19] are embedded in the original design. Hardware obfuscation is usually completed in the functional design of IC [20]. It is designed to prevent the implantation or triggering of HTs by obfuscating the IC function or structural characteristics. Meanwhile, it increases the time, area, and power consumption of the IC design which affects IC performance [21].

Bi et al introduced camouflage technology to create an indistinguishable layout for different circuits [22]. It is designed in the IC layout and wiring by inserting camouflage logic or pseudo connection between the internal layers. It can prevent attackers from using reverse engineering to extract the gate level netlist from the layout image of each layer [23]. However, camouflage design will increase the time cost of IC in the layout and wiring. Besides, the inserted pseudo connection may cause crosstalk and degrades the IC performance [24].

Xiao et al used standard units to fill the blank area of the IC. They can be automatically connected into a combined test circuit and can prevent functional tampering of the standard unit from HT [25]. Its purpose is to fill up the space that can be used to insert HT. Standard units need to be carefully designed by specialized personnel, and this will lead to additional costs [26].

The types of HTs that each active defense can resist are summarized in Tab.2. It can be concluded that all HTs can be resisted by active defenses, but no one active defense can prevent all HT types. The design costs of aforementioned active defense are different and the HT attacks that they can resist are also different. Therefore, it is necessary to quantify the cost and benefits of defense design, and study the game between the attacker and the defender.

TABLE 2. Active defenses and its resisting HT.

Active Defense	HT Resistance
Hardware Obfuscation	Functionality Change
	Information Leakage
Camouflage Technology	Functionality Change
	Performance Degradation
Blank Filling	Information Leakage
	Denial of Service

C. GAME-THEORETIC APPROACHES IN HT DETECTION

To understand the interactions between attackers and testers for HT detection, a number of researches have focused on modeling the interactions using game theory[27-31]. Graf [27] introduced security economic models in conjunction with game theory for guiding system designers in selecting optimal sets of existing HT detection methods.

Subsequently, he presented a game theoretic framework for determining the effectiveness of HT detections. Meanwhile, he illustrated the value of two common solution concepts in HT detection which are the iterated elimination of dominated strategies and Nash equilibrium [28]. Kamhoua et al studied a zero-sum game between HT designer and testers. Multiple possible mixed strategy Nash equilibria is used to identify optimum test sets for increasing the probability of detecting HT [29], [30]. Saad et al proposed a game based on prospect theory to describe the irrational behavior in the strategy selection of both the attacker and defender [31].

The aforementioned game-based methods can be used for HT detection selection, and illustrative numbers are used to verify the methods. The active defenses are also crucial against HT. For the attack-defense confrontation on IC, it provides the possibility to quantify the payoffs of both sides based on hardware resources. We establish the payoffs model for accurately solving Nash equilibrium and select optimal strategy. For HT detection as the last stage of defense, the result of the game between active defenses and the HTs makes the detection more targeted.

III. DYNAMIC GAME BASED ON FUZZY THEORY FOR HT ATTACK AND DEFENSE

A. GAME FRAMEWORK FOR HT ATTACK AND DEFENSE

The active defense designer and the HT designer are the defender and attacker in the game, respectively. Notably, we only discuss the situation that the active defense is designed in the IC design and the HT is inserted in the IC manufacturing in this paper. According to IC design and manufacturing process, the HT is implanted after the active defense design and the attacker can capture the active defense.

Hence, the HT attack-defense problem is formulated as a noncooperative finite dynamic game.

Considering that the complexity and hardware costs of active defense, it is extremely costly to design multiple defenses on an IC. Meanwhile, not all existing ICs have defensive measures. Thus, we assume that the designer, as a defender, adopts no defensive measures or one of the three active defenses which are Hardware obfuscation, camouflage technology and Blank filling. In order to ensure the concealment of the HT, one HT is implanted into an IC at most. The attacker implanted nothing or one of the four HTs in IC manufacturing which are functionality change, performance degradation, information leakage and denial of service.

The purpose of this paper is to understand the interactions between the defender and attacker in IC design and manufacturing. It is necessary to propose a method which can help to understand how the defender and attacker make selections on the type of HTs that they will resist or insert, respectively. This method will provide the balance between attack and defense in IC design and manufacturing. Although the harm of HT cannot be completely eliminated, it rules out several HTs for HT detections after IC packaging. This makes the HT detections more targeted and reduces the cost of it.

For the studied active defenses, the selection of the defender concerning which HTs to resist is impacted by its comprehension of the possible selections of the attacker concerning which type of HTs to insert and vice versa. The selections by both defender and attacker will determine the costs of both sides and the losses faced or avoided by the IC. For this coupling in the actions of the two parties, the noncooperative finite dynamic game theory provides suitable tools for modeling and analyzing which can help to understand the strategy selection processes of the attacker and defender.

We formulate a noncooperative finite dynamic game (NFDG) in strategic form NFDG $= \{N, \{S_i\}_{i\in N}, H, P(h), \{u_i\}_{i\in N}\}$ which is defined by its five main components:

- the players which are the defender d and the attacker a in the set $N = \{d, a\}$,
- the strategy space S_i of each player $i \in N$,
- the sequence of actions from the start of the game to the current decision, H,
- the player function, $P(h)$, which assigns a player to every sequence that is a proper sub history of some terminal history,
- the payoff function u_i of any player $i \in N$.

Obviously, the strategy space of the attacker is the set of studied HT. The attacker can choose one type of HTs to insert or inaction in the IC manufacturing, $\sigma_a \in S_a$. While the strategy space of the defender is the set of studied active defenses. The defender can choose one type of defenses or inaction in the IC design, $\sigma_d \in S_d$.

If the defender resists the wrong type of HT that have been implanted in the IC, denoted by $\sigma_a \neq \sigma_d$. Then, the attacks will cause damage to the IC. This damage is mathematically expressed by a loss L_t while t represents the type of the implanted HT. If the defender resists the right type, denoted by $\sigma_a = \sigma_d$, the IC avoid the loss L_t which can be seen as a defender's benefit. Besides, the designs of active defense and HT are complex and needs to use hardware resources. The costs of attack and defense are mathematically expressed by C_a and C_d, respectively. For each defender's action, its payoff function $u_d(\sigma_d, \sigma_a)$ will be

$$u_d(\sigma_d, \sigma_a) = \begin{cases} L_t - C_d, & if\ \sigma_a = \sigma_d \\ -L_t - C_d, & otherwise \end{cases} \tag{1}$$

For each attacker's action, its payoff function $u_a(\sigma_d, \sigma_a))$ will be:

$$u_a(\sigma_d, \sigma_a) = \begin{cases} -C_a, & if\ \sigma_a = \sigma_d \\ L_t - C_a, & otherwise \end{cases} \tag{2}$$

B. PAYOFF MODELING BASED ON FUZZY VARIABLE

Research of game theory in HT attack and defense has drawn increased attention in recent years. In the payoff calculation, most of the existing researches use illustrative values instead of modeling according to the reality. Considering that payoff analysis is the basis of attack-defense game model and optimal defense strategy selection, this reduces the effectiveness of game theory in practical application. It is worth noting that the intrinsic value and utility of the IC are positively related to its resources. Meanwhile, HT insertion and active defense design both need to occupy IC resources, and their design costs are positively related to the occupied resources. These motivate the need for a payoff quantitative modeling method based on IC resources to study the strategic interactions correctly.

Lee proposed a cost quantification model as the basis of response decision [32]. He compared the response cost with the intrusion loss included operational cost and damage cost. When the response cost is higher than the intrusion loss, the system does not respond. These costs can be qualified according to the attack taxonomy and site-specific security priorities. This model has laid a foundation for network security game. However, its cost quantification idea is oriented to Intrusion detection systems, not to defense decision. Besides, IC security is different from network security in attack types and defense methods.

To solve this problem, we propose a payoff modeling method for HT attack-defense strategy based on Lee's thought as shown in Fig.1. The method quantifies the benefits and costs of attack and defense which contain several factors in the game. Its purpose is to transform the critical factors affecting strategy selection into quantitative values related to cost-benefit and build a general model which contains most critical factors. The results of it provides the expected comprehensive payoff for both entities in the game. Next, we introduce the attack and defense cost-benefit model and its quantitative method.

Attack benefit represents the profits of a successful attack for the attacker. It is generally expressed by the asset loss caused by the attack on IC, L_t, which is a positive value [33]. This benefit is determined by the number of the IC hardware

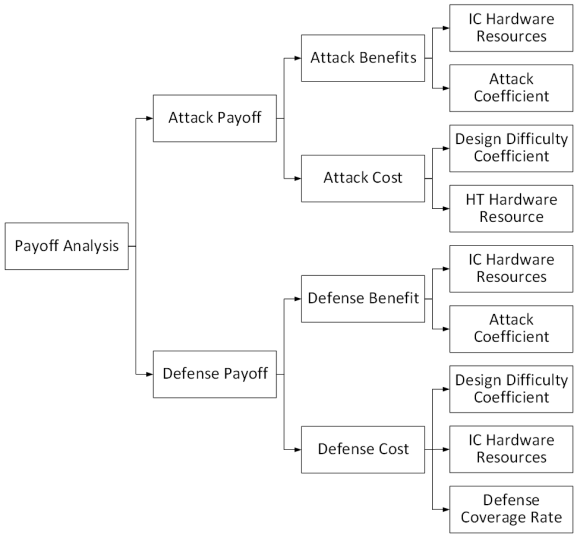

FIGURE 1. Factors analysis of payoff.

resources n_{IC} and the threat severity coefficient K_t of the HT, that is, $L_t = n_{IC} \cdot K_t$. Attacking ICs with more hardware resources can obtain higher profits. And different HTs attack the same chip with different benefits.

Attack cost, C_a, represents the design cost and hardware resources required by an attacker to insert a HT. The design cost is related to the difficulty of implanting the HT. For example, the discovery and utilization of vulnerabilities need time and technical resources. The attack cost is determined by design difficulty coefficient K_a and HT hardware resource cost n_{HT}. The attack cost is calculated as follows.

$$C_a = n_{HT} \cdot K_a \tag{3}$$

After the attack from the denial of service HT, the whole IC cannot be used, and the loss is the largest. The threat severity of it, K_t, is 1. In order to attack the IC, the attacker needs to identify the top module to insert the HT which is difficult in the IC manufacturing. The Design difficulty of it, K_a, is 4. For information leakage HT, the key information of the IC can be leaked, and the important value is stolen. Its K_t, is 0.75. The attacker needs to insert the HT in the memory storing key information and K_a is 3.

The K_t of functionality change is 0.5 for it only changes some functions of the IC, and other functions are still smooth operation. The function can be changed by adding HT into the modules that the function signal flows through and its K_a is 2. The performance degradation HT does not change the function of the IC and its threat severity is the lowest which is 0.25. While the attackers change the doping concentration or wire width in the IC manufacturing which K_a is 3, to cause cumulative damages continuously. The threat severity K_t and design difficulty K_a of each HT is shown in the Tab.3.

It is worth noting that the K_a corresponds to the resources of HT while the K_t corresponds to that of the whole IC. In this case, the K_a is greater than 1 and K_t is less than 1.

In the attack and defense of HT, the ***defense benefit*** is the loss that the IC can avoid after taking the defense design against a certain HT attack. Therefore, the defense benefit is equal to the attack benefit of the successfully resisting HT [10].

TABLE 3. Threat severity coefficient and design difficulty of different HT ty pes.

Category	Threat Severity K_t	Design Difficulty K_a
Denial of Service	1	4
Performance Degradation	0.25	1
Functionality Change	0.5	2
Information Leakage	0.75	3

TABLE 4. Design difficulty of different active defenses.

Category	Design Difficulty K_d
Hardware Obfuscation	0.1
Camouflage Technology	0.2
Blank Filling	0.05

Defense cost represents the defense design cost and hardware resources required by the defender to design active defenses. The defense design cost is related to its difficulty.

Blank filling is the simplest of the three active defenses. It fills the blank space with functional units to remove the space that can be used to insert HT. In general, these units will not affect the original design and its design difficulty, K_d, is the lowest which is 0.05. Hardware Obfuscation is to insert the logic encryption circuit into the original circuit in functional design. The performance impact of the encryption circuit on the whole IC should be considered and its K_d is 0.1. Camouflage Technology is the most complex defense which K_d is 0.2. It adds camouflage logic or pseudo connections between layers in the IC layout which is likely to cause crosstalk and affect IC performance. The design difficulty K_d of each active defense is shown in the Tab.4.

In the attack-defense of HT, as long as the active defense successfully resists the HT, the defender will benefit from it and the payoff is positive. Based on reality, all the K_d is less than the K_t.

The hardware resources used for active defenses are determined by the defense coverage rate, R_d, and the IC resources. Different from the network system attack and defense, we need to pay attention to the defense coverage rate in HT attack and defense. Each defense needs to design multiple active defense units to cover the key functional modules or the whole IC. To achieve high coverage, it is necessary to design more defense units and then more hardware resources are used. Since the defense coverage rate is a quantitative index for HT defense extended from Lee's analysis method,

its impact on defense payoff is uncertain. We use triangular fuzzy variable, $\xi(a, b, c)$ which are commonly used in security analysis, to characterize the impact of defense coverage rate on defense payoff.

$$C_d = n_{IC} \cdot K_d \cdot \xi \cdot R_d \tag{4}$$

C. EXISTENCE PROOF FOR NASH EQUILIBRIUM BASED ON FUZZY THEORY

In the finite dynamic game, the payoff of each strategy is a crisp number, so it is easy to compare the payoff. For noncooperation finite dynamic game based on Fuzzy Theory (NFDG-FT), the payoff of the defender is the function of the fuzzy variable ξ and cannot be compared directly. In view of this situation, we need to modify the general definition of Nash equilibrium.

Let $u = (u_i)_{i\in N}$ be the fuzzy payoff function of any given strategy combination $s^* = (s_i^*)_{i\in N}$. $u_i(s_{-i}^*, s_i^*)$ is the weighted average of the results corresponding to the final historical sequence $(h_1, h_2, \cdots h_m)$. Among them, any final historical sequence h_j is determined by the strategy combination s^*. Set the weight as $\lambda_j \geq 0$, and $\sum_{j=1}^{m} \lambda_j = 1$. $u_i(h_i)$ is the fuzzy payoff of the decision-maker $i \in N$ corresponding to the final historical sequence h_j. Let s_{-i}^* denote a set of strategies of all participants except i. The payoff $u_i(s_{-i}^*, s_i^*)$ is given by

$$u_i(s_{-i}^*, s_i^*) = \lambda_1 u_i(h_1) + \cdots + \lambda_m u_i(h_m) \tag{5}$$

If and only if, under a certain comparison criterion, for any strategy s_i of any decision-maker $i \in N$, $u_i(s_{-i}^*, s_i^*) \geq u_i(s_{-i}^*, s_i)$. We define s_i^* as an equilibrium. For fuzzy variables δ and η, there are three comparison criteria [34].

Definition 1: Expectation criterion. If and only if $E[\delta] < E[\eta]$, $\delta < \eta$. Where the Expectation of δ is defined as

$$E[\delta] = \int_0^{\infty} Cr\{\delta \geq r\}dr - \int_{-\infty}^{0} Cr\{\delta \leq r\}dr \tag{6}$$

Let the Expectation equilibrium be a combination of strategies s^*. For any decision-maker $i \in N$ and its strategy s_i, there are

$$E[u_i(s_{-i}^*, s_i^*)] > E[u_i(s_{-i}^*, s_i)] \tag{7}$$

Definition 2: Optimistic criterion. If and only if $\delta_{\sup}(\alpha) < \eta_{\sup}(\alpha)$ for a given confidence level $\alpha \in (0, 1]$, $\delta < \eta$. Where the α-optimistic value of δ is defined as

$$\delta_{\sup}(\alpha) = \sup\{r |\, Cr\{\delta \geq r\} \geq \alpha\} \tag{8}$$

Let the α-optimistic equilibrium be a combination of strategies s^*. For any decision-maker $i \in N$ and its strategy s_i with a given confidence level $\alpha \in (0, 1]$, there are

$$\begin{aligned}&\sup\{r|Cr\{u_i(s_{-i}^*, s_i^*) \geq r\} \geq \alpha\}\\ &\qquad \geq \sup\{r|Cr\{u_i(s_{-i}^*, s_i) \geq r\} \geq \alpha\}\end{aligned} \tag{9}$$

When a confidence level α is determined, some decision makers in the game tend to avoid risks and maximize optimistic value of his strategy.

Definition 3: Pessimistic criterion. If and only if. $\delta_{\inf}(\alpha) < \eta_{\inf}(\alpha)$ for a given confidence level $\alpha \in (0, 1]$, $\delta < \eta$. Where the α-pessimistic value of δ is defined as

$$\delta_{\inf}(\alpha) = \inf\{r |\, Cr\{\delta \leq r\} \geq \alpha\} \tag{10}$$

Let the α-pessimistic equilibrium be a combination of strategies s^*. For any decision-maker $i \in N$ and its strategy s_i with a given confidence level $\alpha \in (0, 1]$, there are

$$\begin{aligned}&\inf\{r|Cr\{u_i(s_{-i}^*, s_i^*) \leq r\} \geq \alpha\}\\ &\qquad \geq \inf\{r|Cr\{u_i(s_{-i}^*, s_i) \leq r\} \geq \alpha\}\end{aligned} \tag{11}$$

In some cases, the decision maker may be a risk seeker, so the pessimistic value can be used to compare the fuzzy payoff more appropriately. For the attack and defense of HT, the existence of expectation equilibrium is proved as followed.

Theorem 1: Every NFDG-FT for attack and defense of HT has an expected equilibrium in pure strategies.

Let $P(\varphi) = P_1$ and all history sequences with length 1 be $(h_1, h_2, \cdots, h_r - 1, h_r)$. The identification of sub-game is adopted, $(\Gamma(h_1), \Gamma(h_2), \cdots \Gamma(h_r))$. Let $u_i|h$ be the fuzzy payoff of the player $i \in N$. Let $s(j) = (s_{ij})_{i\in N}$be pure strategies for each player in $\Gamma(h_j)(1 \leq j \leq r)$. Then $u_i(s)$,$u_i|h_j(s(j))$ are the payoffs to player i in the NFDG-FT and $\Gamma(h_j)$, respectively.

For the attack and defense of HT, the length of the NFDG-FT is 2. If there is only one player in the game, he can simply choose the strategy which gets the maximized expected fuzzy payoffs. This pure strategy is an expectation equilibrium. The theorem holds for $(\Gamma(h_1), \Gamma(h_2), \cdots \Gamma(h_r))$. Let $s^*(j)$ be the expected equilibrium strategy in $\Gamma(h_j)$. For every strategy s_i of the player i in $\Gamma(h_j)$, there is

$$E[u_i|h_j(s_{-i}^*(j), s_i^*(j))] \geq E[u_i|h_j(s_{-i}^*(j), s_i)] \tag{12}$$

Then, we construct a pure expected equilibrium strategy in NFDG-FT. Let P_1 be a player in N. Without loss of generality, we suppose P_1 to be the defender. We suppose action σ taken by the defender at the initial of the game. when $j = \sigma$ there is

$$\max_{1\leq j\leq r} E[u_i|h_j(s^*(j))] \tag{13}$$

s^* is defined as a pure strategy profile in NFDG-FT.

$$s^* = (s_{-1}^*, s_1^*) \tag{14}$$

where $s^*|h_j = s^*(j)$ and $s^*(\varphi) = \sigma$.

$$\begin{aligned}E[u_1(s_{-1}^*, s_1^*)] &= E[u_1|h_\sigma(s_{-1}^*(\sigma), s_1^*(\sigma))]\\ &\geq E[u_1|h_j(s_{-1}^*(j), s_1^*(j))], \quad 1 \leq j \leq r\end{aligned} \tag{15}$$

For any strategy s_1 of the attacker with $s_1(\varphi) = j$

$$\begin{aligned}E[u_1(s_{-1}^*, s_1)] &= E[u_1|h_j(s_{-1}^*|h_j, s_1|h_j)]\\ &\leq E[u_1|h_j(s_{-1}^*(j), s_1^*(j))]\end{aligned} \tag{16}$$

Since $s^*(j)$ is an expected equilibrium in $\Gamma(h_j)$, for any strategy s_1, there is

$$E[u_1(s_{-1}^*, s_1^*)] > E[u_1(s_{-1}^*, s_1)] \tag{17}$$

For every player $i \in N, i \neq 1$ and any strategy s_i, there is

$$\begin{aligned} &E[u_1(s^*_{-i}, s_i)] \\ &= E[u_i|h_\sigma(s^*_{-i}|h_\sigma, s_i|h_\sigma)] \\ &= E[u_i|h_\sigma(s^*_{-i}(\sigma), s_i|h_\sigma)] \leq E[u_i|h_\sigma(s^*_{-i}(\sigma), s^*_i(\sigma))] \end{aligned} \quad (18)$$

Hence for every player $i \neq 1$ and any strategy s_i, there is

$$E[u_i(s^*_{-i}, s^*_i)] > E[u_i(s^*_{-i}, s_i)] \quad (19)$$

In summary, we have constructed an expected equilibrium s^* in the NFDG-FT. Thus the theorem is shown. The existence of α-optimistic equilibrium and α-pessimistic equilibrium in the NFDG-FT can be proved in the same way.

D. SOLUTION ALGORITHM FOR NASH EQUILIBRIUM

In IC design and manufacture, the HT defense and attack game are divided into two stages. In the first stage, the defender designs an active defense measure in the IC. The attacker inserts the HT in the second stage, and the measure of the defender can be observed by the attacker before he acts. Let S_d be the action space of the defender and S_a be that of the attacker. When the game enters the second stage and the defender chooses $\sigma_d \in S_d$ in the first stage, the problem faced by the attacker is

$$\max_{\sigma_a \in S_a} u_a(\sigma_d, \sigma_a) \quad (20)$$

Obviously, the attacker's optimal choice σ^*_a depends on the choice of the defender σ_d. We use $\sigma^*_a = R_a(\sigma_d)$ to represent the solution of the above optimization problem. While the defender should predict that the attacker's action according to $\sigma^*_a = R_a(\sigma_d)$ in the second stage, the problem faced by the defender is

$$\max_{\sigma_d \in S_d} u_d(\sigma_d, R_a(\sigma_d)) \quad (21)$$

Let the optimal solution of the above problem be σ^*_d. The Nash Equilibrium of this game is $(\sigma^*_d, R_a(\sigma^*_d))$.

To solve the studied HT defense and attack game, we formalized the above analysis and propose a Nash Equilibrium Solution algorithm, summarized in Algorithm 1. In this algorithm, σ_j represents the action of the defender while the action of the attacker is expressed as σ_i. Let σ^*_a be the subgame perfect Nash equilibrium in the second stage and $u_a(\sigma_d, R_a(\sigma_d))$ be the payoff of the attack under the subgame perfect Nash equilibrium. The output of the algorithm is the pure strategy Nash Equilibrium of the game, expressed as $(\sigma^*_d, R_a(\sigma^*_d))$. $u_a(\sigma^*_d, R_a(\sigma^*_d))$ and $u_d(\sigma^*_d, R_a(\sigma^*_d))$ represent the payoff of the attack and the defender under the pure strategy Nash Equilibrium, respectively. The complexity of the algorithm is $O(4n) + O(n^2)$.

IV. CASE STUDY

A. SIMULATION SETUP

According to the hardware resources of HTs in the two researches [35], [36], we applied the proposed method on the HTs of the FPGA in [36]. The FPGA, Kintex-7 XC7K160T-1FBGC, is integrated on the FPGA platform, SAKURA-X board, as the main FPGA. The area overhead of HTs

Algorithm 1 Nash Equilibrium Solution Algorithm

Input: Actions of the defender, m. Actions of the attacker, n.
Output: Pure Strategy Nash Equilibrium of the defender and attacker, $(\sigma^*_d, R_a(\sigma^*_d))$
Initialize $u_a(\sigma_d, R_a(\sigma_d)) = u_a(\sigma^*_d, R_a(\sigma^*_d)) = 0$;
For j=1; j ≤ m+1;j++ **do**
  For i=1;i≤n+1;i++ **do**
    If $u_a(\sigma_j, \sigma_i) > u_a(\sigma_d, R_a(\sigma_d))$ **then**
    $u_a(\sigma_d, R_a(\sigma_d)) = u_a(\sigma_j, \sigma_i)$;
    $R_a(\sigma_d) = \sigma_j$;
    End if
  End for
  If $u_a(\sigma_j, \sigma_a) > u_a(\sigma^*_d, R_a(\sigma^*_d))$ **then**
    $u_a(\sigma^*_d, R_a(\sigma^*_d)) = u_a(\sigma_j, \sigma_a)$;
    $u_d(\sigma^*_d, R_a(\sigma^*_d)) = u_d(\sigma_j, \sigma_a)$;
    $\sigma^*_d = \sigma_j$;
    $R_a(\sigma^*_d) = \sigma_a$;
  End if
End For
Return $(\sigma^*_d, R_a(\sigma^*_d))$.

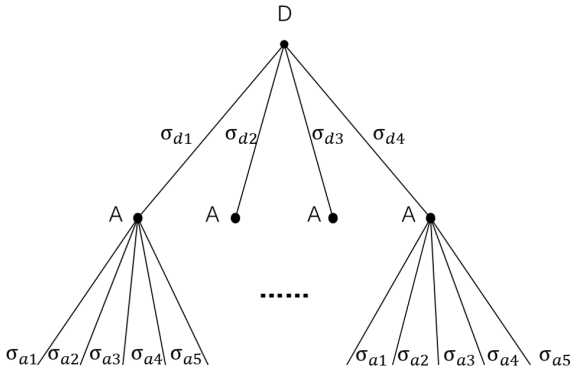

FIGURE 2. Dynamic game for HT attack and defense.

are provided by the number of used LUTs and registers. Considering that some of the HTs may be optimized in the synthesis, the 'Don't Touch' is set for the HT components in Vivado which is the design suite for Xilinx FPGA for synthesis.

For simulating the HT attack and defense game, we consider the scenario in which the defender designs one of the three active defenses into the FPGA or inaction. Let the strategy space S_d = {Hardware obfuscation, Camouflage technology, Blank filling, Inaction} be $S_d = \{\sigma_{d1}, \sigma_{d2}, \sigma_{d3}, \sigma_{d4}\}$. While the attacker inserts one of the four hardware Trojans into the IC or inaction. Let the strategy space S_a = {Functionality Change, Performance Degradation, Information Leakage, Denial of Service, Inaction} be $S_a = \{\sigma_{a1}, \sigma_{a2}, \sigma_{a3}, \sigma_{a4}, \sigma_{a5}\}$, as shown in fig.2. A and D are represent the defender and the attacker, respectively.

In the case study, we first derive the payoff of different strategies for all players based on fuzzy theory. Based on eq.2 and eq.3, for each attacker's action, its payoff function

$u_a(\sigma_d, \sigma_a)$ is

$$u_a(\sigma_d, \sigma_a) = \begin{cases} -n_{HT} \cdot K_a & if\ \sigma_a = \sigma_d \\ n_{IC} \cdot t - n_{HT} \cdot K_a & otherwise \end{cases} \quad (22)$$

We calculate the median resources of each HT type as the HT hardware resource cost, n_{HT}, of the corresponding type based on previous studies and we assume that the number of HT resources is equal to the sum of HT LUTs and Registers in FPGA.

For each defender's action, based on eq.1 and eq.4, its payoff function $u_d(\sigma_d, \sigma_a)$ is

$$u_d(\sigma_d, \sigma_a) = \begin{cases} n_{IC} \cdot t - n_{IC} \cdot K_d \cdot \xi \cdot R_d, & if\ \sigma_a = \sigma_d \\ -n_{IC} \cdot t - n_{IC} \cdot K_d \cdot \xi \cdot R_d, & otherwise \end{cases} \quad (23)$$

We calculate the payoff of different defender's actions when the defense coverage, R_d, is 100%.

Then, we use Algorithm 1 to solve the Nash equilibrium under three different criteria which allow us to gain more insights on the proposed game.

B. PAYOFF ANALYSIS AND EQUILIBRIUM SOLUTION

According to the hardware resources of HTs in previous studies, there are some extreme data in Functionality Change, Denial of Service and Information Leakage. We calculate the median resources of each HT type as the HT hardware resource cost of the corresponding type. The HT hardware resource cost of different types, n_{HT}, is shown in Tab.5. In particular, we assume that the number of HT resources is equal to the sum of HT LUTs and Registers in FPGA.

As shown in Tab.5, n_{HT} of σ_{a1},σ_{a2},σ_{a3},σ_{a4} are 44.5, 59, 196.5, 40, respectively. For,σ_{a5}, the attacker does not change the FPGA while there are no costs and benefits. According to the eq.21, we calculate the payoff of different attacker's actions which are shown in Tab.6.

From the results of attack payoff analysis, it can be concluded that the cost of action σ_{a2} is the highest and that of action σ_{a3} is the lowest. When the HT attacks the IC successfully, the payoff of action σ_{a4} is the highest and that of action σ_{a3} is the lowest. While the HT is successfully defended by the active defense, its payoff is negative. Even so, the attackers are preferable to insert a HT in the IC whcih is consistent with the actual situation. For the successful attack of HT, the benefit is far more than the cost. Besides, considering that the attacker can observe the defender's action, the attacker will avoid the defensive strategy of the defender.

According to the intermediate results of the algorithm 1, the Nash equilibrium of subgame is (σ_{d1},σ_{a4}), (σ_{d2},σ_{a4}), (σ_{d3},σ_{a1}), (σ_{d4},σ_{a4}) which are the optimal strategies of the attacker for each action of the defender.

For payoff function of defender's action, ξ is a fuzzy variable. Let $u_d(\sigma_d, \sigma_a) = (u_{exp}, u_{opt}, u_{pes})$ be the expected, optimistic and pessimistic value of defense action payoff, respectively. We assume that confidence level α_{opt},α_{pes} are 0.8 and 0.6, respectively. According to the eq.22, when the defense coverage is 100%, we calculate the payoff of different defender's actions which are shown in Tab.7.

From the results of defense payoff analysis, it can be obtained that the cost of action σ_{d2} is the highest and that of action σ_{d3} is the lowest. When the defender doesn't design

TABLE 5. Resources of different HTs.

Category	Benchmarks	Num of HT resources	Median of HT resources
Functionality Change	RS232-T1000	44	44.5
	RS232-T1100	44	
	RS232-T1200	45	
	RS232-T1300	31	
	RS232-T1400	50	
	RS232-T1500	48	
	RS232-T1600	39	
	S35932-T100	34	
	S38584-T200	198	
	S38584-T300	976	
Performance Degradation	S35932-T300	59	59
Information Leakage	AES-T100	89	196.5
	AES-T200	89	
	AES-T300	64	
	AES-T400	314	
	AES-T600	400	
	AES-T700	90	
	AES-T800	94	
	AES-T900	383	
	AES-T1000	90	
	AES-T1100	94	
	AES-T1200	374	
	AES-T1300	65	
	AES-T1400	69	
	AES-T1500	359	
	AES-T1600	317	
	AES-T1700	471	
	AES-T2000	299	
	AES-T2100	686	
Denial of Service	AES-T500	133	40
	AES-T1800	129	
	AES-T1900	420	
	S15850-T100	61	
	S35932-T200	40	
	S38417-T100	29	
	S38417-T200	35	
	S38417-T300	31	
	S38584-T100	21	

TABLE 6. Payoff of different attacker's actions.

	σ_{d1}	σ_{d2}	σ_{d3}	σ_{d4}
σ_{a1}	-89.0	-89.0	2168.0	2168.0
σ_{a2}	1069.5	-59.0	1069.5	1069.5
σ_{a3}	-589.5	2796.0	-589.5	2796.0
σ_{a4}	4354.0	4354.0	-160.0	4354.0
σ_{a5}	0	0	0	0

TABLE 7. Payoff of different defender's actions.

	σ_{d1}	σ_{d2}	σ_{d3}	σ_{d4}
σ_{a1}	1805.6	1354.2	-2482.7	-2257.0
	1859.7	1462.5	-2455.6	-2257.0
	1787.5	1318.0	-2491.7	-2257.0
σ_{a2}	2934.1	-4288.3	3159.8	-3385.5
	2988.2	-4179.9	3186.8	-3385.5
	2916.0	-4324.4	3150.7	-3385.5
σ_{a3}	-1579.9	225.7	-1354.2	-1128.5
	-1525.7	334.0	-1327.1	-1128.5
	-1597.9	189.5	-1363.2	-1128.5
σ_{a4}	-4965.4	-5416.8	4288.3	-4514.0
	-4911.2	-5308.4	4315.3	-4514.0
	-4983.4	-5452.9	4279.2	-4514.0
σ_{a5}	-451.4	-902.8	-225.7	0
	-397.2	-794.464	-198.6	0
	-469.4	-938.912	-234.7	0

any active defense, expressed as σ_{d4}, its payoffs are all non-positive. While the attack is successfully defended, the benefit is much higher than the cost. Therefore, the defender is more inclined to design active defense against HT. Although the attacker can learn about the defender's action, the defender is willing to design active defense to resist the threat of some HTs whcih is consistent with the actual situation. In addition, considering the maximum defense cost, that is, 100% defense coverage, fuzzy variables have little effect on strategy selections. There is an order of magnitude gap between defense cost and defense benefit.

According to the result of the sub game Nash equilibrium, the Pure Strategy Nash Equilibrium of the game is $(\sigma_{d3}, \sigma_{a1})$ under the three criteria. In this strategy, the payoffs of the attacker and defender are 2168.0 as shown in Tab.6 and -2482.7 as shown in Tab.7, respectively. The attacker can know the action of the defender and HT will not be successfully defended. It is worth noting that the purpose of the method for the defender is to avoid the threat of some HTs at the lowest cost. Although the harm of HT cannot be completely eliminated in the design and manufacturing stage, it rules out several HT kinds for HT detections after IC packaging. This make the HT detections more targeted and reduces the cost of it.

Compared with the related works, the proposed resource-based payoff model is more practical than the illustrative numbers. The modeling takes many critical factors into account in HT attack and active defense. Meanwhile, all the actions of the both sides correspond to the actual HT attack and active defense technology. Thus, it can be used to guide the design of HT active defenses. The solving algorithm is proposed to obtain the Nash equilibrium of the game with the complexity of $O(4n) + O(n^2)$. It is less than the complexity of the algorithm in [31] which is $O(10n) + O(n^2)$. This means that the proposed algorithm takes less time to obtain the Nash equilibrium automatically.

The proposed resource-based payoff model considering design difficulty, attack effect and defense coverage is more practical than the illustrative numbers of the existing methods. Triangular fuzzy number is used to characterize the effect of defense coverage rate on defense payoff. Attack and defense strategies correspond to hardware Trojan types and active defense measures in reality, respectively. Meanwhile, through the analysis of the results, it is found that the results are consistent with the actual situation. All these ensure the effectiveness of the proposed method.

V. CONCLUSION

In this paper, a dynamic game method for modeling the interactions between the attacker, who inserts HT in IC manufacturing, and the defender, who adds active defense in IC design. The problem of HT attack and defense is formulated as noncooperative dynamic game. The attacker chooses the optimal HT type to insert after knowing the defender's active defense. The payoff models of the attacker and defender are proposed based on the IC hardware resources to solve the Nash equilibrium accurately. To account for the uncertainty of the defense coverage on the payoff of defender, a fuzzy variable is used to represent the uncertainty.

Subsequently, the existence of Nash equilibrium in the game with fuzzy payoff is proved. A solution algorithm for pure strategy Nash Equilibrium is proposed to alleviate the threat of some HTs at the lowest cost for the defender. Through the case study of FPGA based on the known HT types and active actions, the use of the fuzzy-theoretic considerations can provide the optimal strategies of the attacker and defender. The result shows that the best optimal strategy is the defender designs blank filling and the attacker inserts functionality change HT. The payoff of the attacker and defender are 2168.0 and - 2482.7, respectively. Although the HT is inserted and avoid the active defense successfully, it can be foreseen and resisted by HT detection.

REFERENCES

[1] T. Schultz, R. Jha, M. Casto, and B. Dupaix, "Vulnerabilities and reliability of reRAM based pufs and memory logic," *IEEE Trans. Rel.*, vol. 62, no. 2, pp. 690–698, Jun. 2019.
[2] Z. Huang, Q. Wang, Y. Chen, and X. Jiang, "A survey on machine learning against hardware trojan attacks: Recent advances and challenges," *IEEE Access*, vol. 8, pp. 10796–10826, 01 2020.
[3] S. Bhunia and M. Tehranipoor, *Introduction to Hardware Security*. San Francisco, CA, USA: Morgan Kaufmann, 2019, pp. 1–20.
[4] J. Wang, S. Guo, Z. Chen, and T. Zhang, "A benchmark suite of hardware trojans for on-chip networks," *IEEE Access*, vol. 7, pp. 102002–102009, 07 2019.
[5] S. Gokulanathan, L. Srivani, D. Thirugnana Murthy, K. Madhusoodanan, and S. Murty, "A review on ht attacks in pld and asic designs with potential defence solutions," *IETE Tech. Rev.*, vol. 15, pp. 1–14, Jan. 2017.
[6] C. Dong, Y. Liu, J. Chen, X. Liu, W. Guo, and Y. Chen, "An unsupervised detection approach for hardware trojans," *IEEE Access*, vol. 8, pp. 158169–158183, 07 2020.
[7] S. Bhunia, M. S. Hsiao, M. Banga, and S. Narasimhan, "Hardware trojan attacks: Threat analysis and countermeasures," *Proc. IEEE*, vol. 102, no. 8, pp. 1229–1247, Aug. 2014.
[8] R. R. Brooks, J.-E. Pang, and C. Griffin, "Game and information theory analysis of electronic countermeasures in pursuit-evasion games," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 38, no. 6, pp. 1281–1294, Nov. 2008.

[9] Q. Meng, S. Tan, Z. Li, B. Chen, and W. Shi, "A review of game theory application research in safety management," *IEEE Access*, vol. 8, pp. 107301–107313, 2020.
[10] S. Kim, "Game theory for network security," in *Proc. Res. Pract.* Jun. 2017, pp. 369–382.
[11] K. S. Kumar, R. Chanamala, S. R. Sahoo, and K. K. Mahapatra, "An improved AES hardware trojan benchmark to validate trojan detection schemes in an ASIC design flow," in *Proc. 19th Int. Symp. VLSI Des. Test*, Jun. 2015, pp. 1–6.
[12] B. Shakya, T. He, H. Salmani, D. Forte, S. Bhunia, and M. Tehranipoor, "Benchmarking of hardware trojans and maliciously affected circuits," *J. Hardw. Syst. Secur.*, vol. 1, no. 1, pp. 85–102, Mar. 2017.
[13] K. Yang, M. Hicks, Q. Dong, T. Austin, and D. Sylvester, "A2: Analog malicious hardware," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 18–37.
[14] K. S. Subramani, N. Helal, A. Antonopoulos, A. Nosratinia, and Y. Makris, "Amplitude-modulating Analog/RF hardware trojans in wireless networks: Risks and remedies," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3497–3510, Apr. 2020.
[15] J. A. Roy, F. Koushanfar, and I. L. Markov, "Ending piracy of integrated circuits," *Computer*, vol. 43, no. 10, pp. 30–38, Oct. 2010.
[16] G. S. Rose, "A chaos-based arithmetic logic unit and implications for obfuscation," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Jul. 2014, pp. 54–58.
[17] M. S. Samimi, E. Aerabi, Z. Kazemi, M. Fazeli, and A. Patooghy, "Hardware enlightening: No where to hide your hardware trojans!" in *Proc. IEEE 22nd Int. Symp. On-Line Test. Robust Syst. Des. (IOLTS)*, Jul. 2016, pp. 251–256.
[18] P. Subramanyan, S. Ray, and S. Malik, "Evaluating the security of logic encryption algorithms," in *Proc. IEEE Int. Symp. Hardw. Oriented Secur. Trust (HOST)*, May 2015, pp. 137–143.
[19] B. Khaleghi, A. Ahari, H. Asadi, and S. Bayat-Sarmadi, "FPGA-based protection scheme against hardware trojan horse insertion using dummy logic," *IEEE Embedded Syst. Lett.*, vol. 7, no. 2, pp. 46–50, Jun. 2015.
[20] C. Dunbar and G. Qu, "Designing trusted embedded systems from finite state machines," *ACM Trans. Embedded Comput. Syst.*, vol. 13, no. 5s, pp. 1–20, Dec. 2014.
[21] B. Liu and B. Wang, "Embedded reconfigurable logic for ASIC design obfuscation against supply chain attacks," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, 2014, pp. 1–6.
[22] Y. Bi, P.-E. Gaillardon, X. S. Hu, M. Niemier, J.-S. Yuan, and Y. Jin, "Leveraging emerging technology for hardware Security–Case study on silicon nanowire FETs and graphene SymFETs," in *Proc. IEEE 23rd Asian Test Symp.*, Nov. 2014, pp. 342–347.
[23] R. Cocchi, J. Baukus, L. Chow, and B. Wang, "Circuit camouflage integration for hardware ip protection," in *Proc. Des. Autom. Conf.*, Jun. 2014, pp. 1–5.
[24] J. Rajendran, M. Sam, O. Sinanoglu, and R. Karri, "Security analysis of integrated circuit camouflaging," in *Proc. Conf. Comput. Commun. Secur.*, Nov. 2013, pp. 709–720.
[25] K. Xiao, D. Forte, and M. Tehranipoor, "A novel built-in self-authentication technique to prevent inserting hardware trojans," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 12, pp. 1778–1791, Dec. 2014.
[26] P.-S. Ba, S. Dupuis, M. Palanichamy, M.-L. Flottes, G. D. Natale, and B. Rouzeyre, "Hardware trust through layout filling: A hardware trojan prevention technique," in *Proc. IEEE Comput. Soc. Annu. Symp.*, Jul. 2016, pp. 254–259.
[27] J. Graf, "Toward optimal hardware trojan detection through security economics and game theory," in *Proc. GOMACTech*, Orlando, FL, USA, Mar. 2016, p. 3.
[28] J. Graf, "Trust games: How game theory can guide the development of hardware trojan detection methods," in *Proc. IEEE Int. Symp. Hardw. Oriented Secur. Trust*, May 2016, pp. 91–96.
[29] C. A. Kamhoua, M. Rodriguez, and K. A. Kwiat, "Testing for hardware trojans: A game-theoretic approach," in *Decision and Game Theory for Security* (Lecture Notes in Computer Science), vol. 8840. Cham, Switzerland: Springer, 2014, pp. 360–369.
[30] C. A. Kamhoua, H. Zhao, M. Rodriguez, and K. A. Kwiat, "A game-theoretic approach for testing for hardware trojans," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 2, no. 3, pp. 199–210, Jul. 2016.
[31] W. Saad, A. Sanjab, Y. Wang, C. A. Kamhoua, and K. A. Kwiat, "Hardware trojan detection game: A prospect-theoretic approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7697–7710, Sep. 2017.
[32] W. Lee, W. Fan, M. Miller, S. J. Stolfo, and E. Zadok, "Toward cost-sensitive modeling for intrusion detection and response," *J. Comput. Secur.*, vol. 10, nos. 1–2, pp. 5–22, Jan. 2002.
[33] S. Hasan, A. Dubey, G. Karsai, and X. Koutsoukos, "A game-theoretic approach for power systems defense against dynamic cyber-attacks," *Int. J. Electr. Power Energy Syst.*, vol. 115, Feb. 2020, Art. no. 105432.
[34] B. Liu, *Theory and Practice of Uncertain Programming*, vol. 239. Berlin, Germany: Springer-Verlag, Jan. 2009.
[35] C. Dong, F. Zhang, X. Liu, X. Huang, W. Guo, and Y. Yang, "A locating method for multi-purposes HTs based on the boundary network," *IEEE Access*, vol. 7, pp. 110936–110950, 2019.
[36] M. Xue, R. Bian, J. Wang, and W. Liu, "Building an accurate hardware trojan detection technique from inaccurate simulation models and unlabelled ICs," *IET Comput. Digit. Techn.*, vol. 13, no. 4, pp. 348–359, Jul. 2019.

DAMING YANG received the B.S. degree in quality and reliability engineering from the School of Reliability and Systems Engineering, Beihang University, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree in advanced manufacturing technology. His research interests include hardware Trojan design and detection.

CHENG GAO was born in Aohan Banner, Chifeng, Inner Mongolia, China, in 1970. He received the B.S. degree in electronic engineering and the M.S. degree in signal and information systems from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 1994 and 2001, respectively, and the Ph.D. degree in aerospace systems engineer from the School of Reliability and Systems Engineering, Beihang University, in 2011. He is currently a Professor and the Director of the Research Center for Component Quality Engineering, Beihang University. His research interests include testing and reliability evaluation of LSIC. He was a recipient of the first prize of Chinese Mechanical Engineering Scientific and Technological Progress.

JIAOYING HUANG received the B.S. degree in thermal energy and power engineering from the Department of Mechanical Engineering, Hu'nan University, Changsha, Hu'nan, China, in 2000, and the M.S. degree in theory and new technology of electrical engineering and the Ph.D. degree in electrical engineering from the Department of Electrical Engineering, Hu'nan University, in 2002 and 2008, respectively. From 2003 to 2004, she was an Assistant Engineer with the Institute of Automation Chinese Academy of Sciences. From 2008 to 2010, she was a Postdoctoral Fellow with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. She is currently a Senior Engineer and the Vice Director of the Research Center for Component Quality Engineering, Beihang University. Her research interests include test, analysis, and evaluation of reliability of electronic components.