A Feature Extraction Method for Hardware Trojans Detection

Zhao Zhixun [1, a], Ni lin [2, b], Li Shaoqing [1, c] and Shi Yubo[1, d]

[1] College of Computer, National University of Defense Technology, 410073, Changsha, China

[2] Xi'an Communication College, 710106, Xi'an, China

[a]zhaozhixun1991@163.com

Keywords: Hardware Trojans; Feature extraction; Convergence and Divergence Analysis

Abstract. Recently Hardware Trojan has been widely studied for its threat to IC security. The method based on power side-channel information is an effective one in various Hardware Trojan detection methods proposed at present. However, there is a problem that Hardware Trojan power is difficult to be distinguished and easy to be drowned by the noise if using the power information to recognize Hardware Trojan through direct power difference during detection process. On the base of power component analysis, this paper proposes that we can extract the feature character of power information based on the feature difference between Hardware Trojan power and noise to manifest the power caused by Hardware Trojan. Then it gives out the power feature extraction algorithm and Trojan power identification model. In the experiment, the AES circuit is used as attack carrier embedded into the hardware Trojan circuit, then simulating the power, the results show that the feature extraction algorithm and convergence and divergence identification model can detect Hardware Trojans effectively.

1 Introduction

With the development of integrated circuits, integrated circuits become more and more complex in functionality and scale. To reduce the cost and chip production cycle, most of the IC design company use wide range of third-party IP cores to accelerate the design. Meanwhile, the design and manufacturing process is separated, the manufacturing process is usually did by chip foundries. As this process which can't be controlled by designers join in, the safety and reliability of chips become a great challenge in the view of quality systems. Hardware Trojans which was widespread concerned in recent years is urgent to research [1, 2, 3].

In essence, Hardware Trojans is additional logic which insert in circuits by redundant parts of the original circuit. It will affect the normal function or leaking work information of chips. Hardware Trojans circuit is triggered by monitoring chip works or accepting external control signal. When the trigger conditions are met, Trojan circuit will execute functions preset by attacker such as information stealing, side-channel leakage, accelerating invalidation, logic destruction, physical destruction, resources occupation and so on [1,2]. Hardware Trojans are usually divided into two parts, the trigger logic unit and functional logic unit. The trigger logic unit responsible for monitoring the external input signal, when the trigger signal arrives, the functional logic unit will be activated. The functional logic is the main part of Hardware Trojans, it will change the logic status of original circuit or cause other Trojan attacks. For internal signal triggered Hardware Trojan, attackers usually use rarely appeared node status or various trigger conditions in design to prevent spurious triggering and improve concealment of Hardware Trojans. Due to the diversity of functions and circuit structures, the designers have many choices when they design a functional logic unit. So it's difficult to present a typical Hardware Trojans structure in circuit, and it increase the difficulty of detection [4].

2 Hardware Trojans detection

Currently, Hardware Trojans detection has not formed a certain industry technical standards, most detection studies are at the laboratory stage. It's difficult to completely detect Hardware Trojans for a number of chips. Now, several big changes of Hardware Trojan detection is:

1. Lack observability and controllability in the process of chip production. 2. The scale increasing and extensive use of third-party IP cores make the region in which Hardware Trojans could easily hide become larger. It's very difficult to complete detections to all the IP cores and circuit module. 3. The detection method based on reverse anatomy is highly reliable, but it's a kind of destructive detection. It can only be used in small portion of the batch chips. The detection completeness is not high enough. 4. Hardware Trojan is usually well hidden and triggered subtle. It's difficult to activate and detect Hardware Trojans. 5. With the development of the semiconductor industry, manufacturing environment and technological factors have a significant impact on the detection process.

Existing non-destructive hardware Trojan detection research main focus on the following aspects: 1. Try to improve reversal rate and improve the production rate of rare state by producing detection code. Identify abnormal circuit by suitable detection; 2. Monitor circuit work status through specialized sensors and additional circuit design, to detect Hardware Trojans by the self-test of measurement circuit. The advantage of this method is that GOLDEN IC is not needed as comparison. 3. Measure the bypass information of circuit when chip works. To identify is there any abnormality in circuit by contrasting the bypass information characteristic fingerprint database and appropriate recognition model.

Hardware Trojans detection method proposed in this paper is based on the measurement of power information. In the circuit operation, if the Hardware Trojan is activated, certainly the power consumption changes. In normal case, the dynamic power consumption of chips is described as : $P_{dynamic} = CLV^2 f_{0\to1}$. Considering the influence of noise and Hardware Trojans, the total power consumption of chips is described as : $P_{chip} = P_{dynamic} + e(pvt) + e(trojans)$. $e(pvt)$ is noise introduced by the circuit itself, it's usually caused by process variation, working voltage and temperature; $e(trojans)$ is power influence caused by Hardware Trojans. Under the premise to eliminate the influence of noise, the most direct way of Hardware Trojans detection is make difference to the power consumption of original circuit and circuit with Hardware Trojans. However, the power consumption of circuit with Hardware Trojans is usually small. And it's difficult to completely eliminate power consumption of noise, which may cover the power consumption caused by Hardware Trojans. Therefore, direct difference method is not feasible.

So a major bottleneck of Hardware Trojans detection is how to make the power consumption caused by Hardware Trojans more obvious. During the detection, we can carry out several tests at the same voltage and temperature to eliminate the effects of voltage and temperature. In the actual chip testing, the process variations is stable existed and have a great impact on elimination of noise. So the focus of the research is how to reduce their impact on power consumption caused by Hardware Trojans during the clear process. If the Hardware Trojans is activated, the power consumption caused by Hardware Trojans must influence the variance of P_{chip} , and this change is not available to process variation. Therefore, if we can enlarge the difference of power consumption variance, then we can enlarge the influence of Hardware Trojans. So our problems on the Hardware Trojans detection attribute to how to enlarge the variance of power consumption matrix.

For number n samples, we can get the original data matrix of power consumption by observing number p sampling points for each sample.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \underline{\underline{\Delta}} (X_1, X_2, \cdots, X_p)$$

X1,…,Xp is vectors of X, and their linear combination is:

$$F = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \underline{\underline{\Delta}} a'X$$

$a = (a_1, a_2, \cdots, a_p)'$, $X = (X_1, X_2, \cdots, X_p)'$. The characteristic root of origin power consumption's covariance matrix Σ is $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, corresponding unit eigenvector is $u_1, u_2, \cdots, u_p$.

$$a'\Sigma a = \sum_{i=1}^{p} \lambda_i a' u_i u_i' a = \sum_{i=1}^{p} \lambda_i (a'u_i)(a'u_i)' = \sum_{i=1}^{p} \lambda_i (a'u_i)^2$$

When $a = u_1$:

$$u_1'\Sigma u_1 = u_1' \left(\sum_{i=1}^{p} \lambda_i u_i u_i' \right) u_1 = \sum_{i=1}^{p} \lambda_i u_1' u_i u_i' u_1 = \lambda_1 (u_1' u_1)^2 = \lambda_1$$

So $a = u_1$ makes $Var(a'X) = a'\Sigma a$ maximum, and

$$Var(u_1'X) = u_1'\Sigma u_1 = \lambda_1$$

As the same reason: $Var(u_i'X) = \lambda_i$, $Cov(u_i'X, u_j'X) = \sum_{a=1}^{p} \lambda_a (u_i'u_a)(u_a'u_j) = 0, i \neq j$

Above derivation shows that the main component of $X_1, X_2, \cdots, X_p$ is a linear combination whose coefficient is eigenvector of Σ . $X_1, X_2, \cdots, X_p$ is unrelated, and it's variance is the characteristic root of Σ .

After feature extraction described above, we can get two sets of data in power matrix projection space. The original circuit X1 and circuit under test X2, X1=（x11,…,x1n）and X2=（x21,…,x2n）. Suppose that X1 obey distribution of F , there is a assumption according to the theory of probability and statistics.

$$H_0 : F \in \Gamma_0$$

X1 obey distribution of Γ_0 . Opposite assumption is:

$$H_1 : F \notin \Gamma_0 \text{ or } H_1 : F \in \Gamma_1, \Gamma_0 \bigcap \Gamma_1 = \varnothing$$

Γ_1 is the distribution of F . To test this assumption, the common way is to confirm the difference between Γ_0 and Γ_1 , described as $m(\Gamma_0, \Gamma_1)$. And $m(\Gamma_0, \Gamma_1)$ has to satisfy the following conditions:

(1) $m(\Gamma_0, \Gamma_1) = 0 \Leftrightarrow \Gamma_0 \equiv \Gamma_1$;

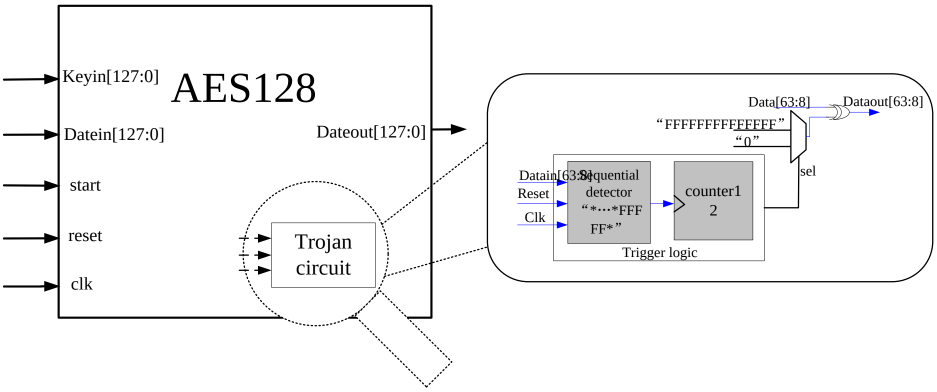
(2) $m(\Gamma_0, \Gamma_1) \geq 0$, and the greater its value, the more difference between Γ_0 and Γ_1 .

So when the value of $m(\Gamma_0, \Gamma_1)$ is small, H_0 is accepted, otherwise, H_0 is not accepted.

3 Experiments analysis based on AES

3.1 The design and implementation of Hardware Trojans circuit

The AES algorithm circuit is used as an attack target in experiment. A type of function tamper Hardware Trojans circuit is designed and implemented based on 56 bit sequential detector and 4 bit counter in NC-verilog. The structure of Hardware Trojans in AES circuit is showed in Figure 3.1. Hardware Trojans circuit trigger the counter by monitoring the specific sequence of external input. Hardware Trojans circuit is activated when the counter reaches a certain value, and changes the result of output.



Piture3.1: The design and implementation of Hardware Trojans circuit

For the design showed above, AES circuit and Hardware Trojans circuit is synthesized in 40nm standard cell library by Synopsys Design Compiler. And then, Hardware Trojans is insert in network. Information of the circuit area is listed in table 3.1:

Table 3.1: Information of Hardware Trojans circuit area

	Number of PMOS	Number of NMOS	area（um^2）
Original circuit	31864	31864	13512
Hardware Trojans circuit	539	539	228.378
Percentage of increase	1.69%	1.69%	1.69%

The result shows that the total area of target circuit is 13512.0048 um^2. Because of CMOS process library, the number of nmos and pmos both are 31864; the area of Hardware Trojans circuit is 228.378 um^2, and proportion is 1.69%.The number of nmos and pmos both are 539, and proportion is 1.69%.

3.2 Analysis of experimental results

Using test vectors as plaintext input, extract transistor-level net list and simulate it in HSPICE. After simulation, obtain and analyze the data of power consumption, then accomplish Hardware Trojan detection.

For each under test chip, obtain the power consumption and calculate the mean value C_i. Power consumption of under test chips are C=（C_1,C_2,……,C_n）. Then we can get the mean power consumption C_{mean}. We can get the centralize power consumption matrix $S=C-C_{mean}$. Caculate characteristics manifest matrix and subspace matrix of power consumption of under test chips, described as K and S_T. After that we can get the subspace matrix of power consumption of contrast chips described as S_G in the same way[6].

During the process of detection, implement eigentransformation of AES circuit and UTC. Extract the first three dimensionality data of S_G and S_T as subspace matrix of power consumption. Obtain the projection data in subspace with the distance between original point and analyze the projection data. There are 20 power consumption curves which reflecting the process variation, we can get a $20 \times n$ characteristic matrix after conversion process.

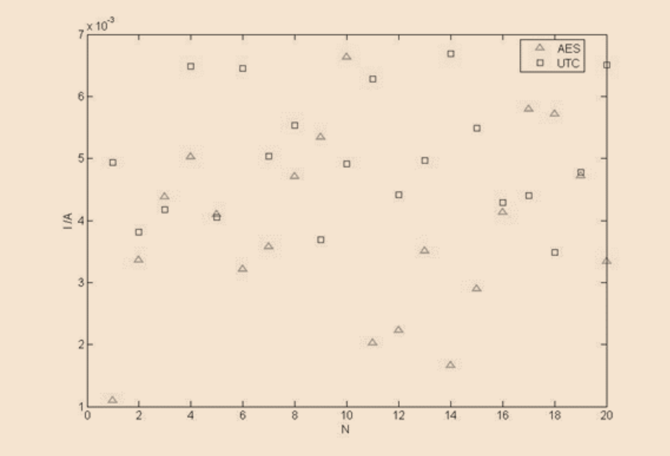


Figure3.2: Projective distribution of power consumption data

Projective distribution of power consumption data is showed in Figure 3.2. Red dot indicates the power consumption data of under test circuit with Hardware Trojans while green dot indicates projective power consumption of original AES circuit. As it showed in the figure, projective power consumption of under test circuit with Hardware Trojans is obviously higher than original AES circuit. The difference and total distribution of projective power consumption is not the same in different point. It shows that the projective power consumption difference can be used as a sign of Hardware Trojans extraction.

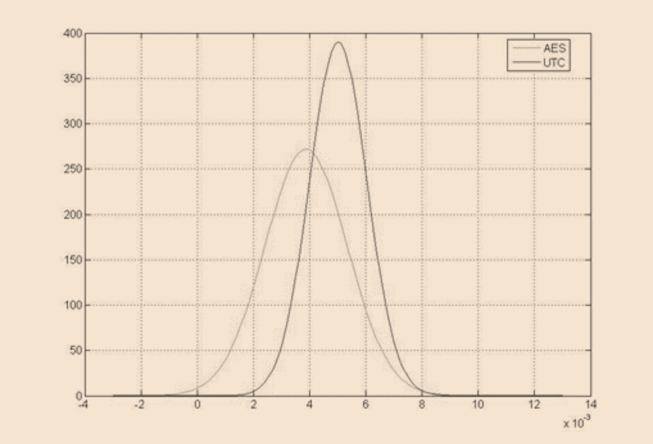


Figure3.3: Normal distribution of projective power consumption

Observe the value characters after finishing mathematical statistics to original AES projective power consumption data. As figure3.3 shows, projective power consumption expectation of under test circuit with Hardware Trojans is obviously higher than original AES circuit. During the projection process, power consumption matrixes of under test circuit and original AES circuit both multiply by the same feature projection matrix, so the relative power difference do not change. Meanwhile, in mathematical statistics, we can also find that power consumption data have difference in distribution. Projective power consumption have various deviation from the expectation. To show the deviation in a better way, we obtained a convergent sequence whose convergence value is projective power expectation.

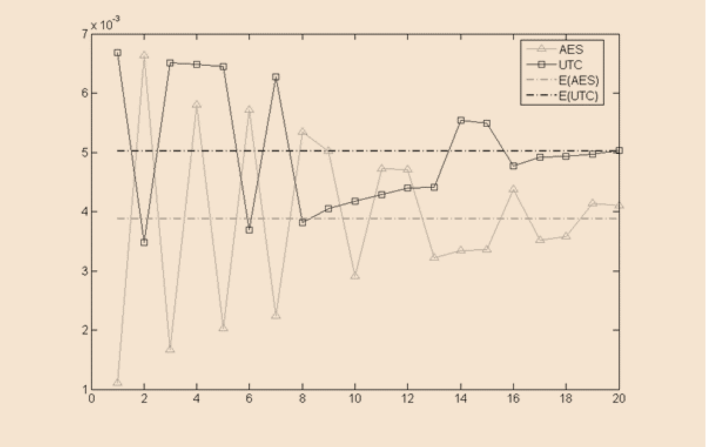


Figure3.4: Distribution trends of data convergence

As is shown in figure3.4, projective value is converge from expect value, and the power consumption of under test circuit converge faster than original AES circuit. In addition, it’s deviation is smaller than original circuit. From the whole feature extraction, figure transformation of matrix substantially is to manifest the variance of power consumption. Due to existing of Hardware Trojans power, distribution of projective power consumption is more concentrate. Thus, the distribution of convergence is increased by projection. Through this character, we can distinguish the difference between under test circuit and original circuit in power consumption and detect Hardware Trojan.

4 Summary

For the process variation noise issue in power consumption information extraction, this paper proposed a Hardware Trojans extraction method that by feature extraction of power consumption data and analyze the distribution of projective power consumption in projection space. Through the experiments based on AES circuit, proved that the method based on figure projection have obvious effect in optimizing process variation noise and the contract recognition model in figure projection space is effective. It provides a feasible and effective solution for the noise detection and algorithm issues in power consumption bypass information Hardware Trojans detection method.

References

[1] Nilin, Lishaoqing, Maruicong, Weipei. Hardware Trojans Detection and Protection. [J].Digital Communication, 2014, 41(1) :59~63.

[2] Agrawal D, Baktir S, Karakoyunlu D, et al. Trojan detection using IC fingerprinting[C]. Proceeding of the 2007 IEEE Symposium Security and Privacy. Oakland:CA, USA. 2007:296-310.

[3] NI Lin, LI Shaoqing, Chen Jihua, Wei Pei, Zhao Zhixun. The influence on sensitivity of Hardware Trojans detection by test vector [C] //proceeding of 2014 communications security conference（CSC2014）. Beijing, china, 2014. pp. 46-51.

[4] Suiqiang, Hardware Trojans Detection Based on Sidechannel Signal Analysis[D], National University of Defense Technology, 2012.

[5] Chakraborty R S，Narasimhan S，Bhunia S. Hardware Trojan: Threats and emerging solutions［C］//Proceedings of High Level Design Validation and Test Workshop (HLDVT’2010). 2010: 166-171.

[6] Zhaozhixun，Nilin，Lishaoqing. Hardware Trojans detection based on dynamic current analysis. The 21st China Academic Annual Conference On Information Theory，2014，pp:7