Thermal maps based HT detection using spatial projection transformation

Yongkang Tang1, Shaoqing Li1, Fan Zhang2, Liang Fang1

1School of Computer, Institute of Microelectronics and Microprocessors, National University of Defense Technology, Sanyit Street, Kafu District, Changsha, People’s Republic of China
2College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, People’s Republic of China
✉ E-mail: tangyongkang17@nudt.edu.cn

Abstract: Hardware Trojan (HT) is increasingly becoming a serious problem in the information security field. Compared to other countermeasures, thermal maps based detection can mitigate process variation (PV) and have a higher accuracy. However, HT cannot be differentiated from the others directly from the original thermal maps. Therefore, in this study, the authors first propose a general HT detection framework based on difference temperature matrix, and introduce the PV mitigation mechanism. Then, they demonstrate how principal component analysis can implement spatial projection transformation and expose HT signal. Finally, they introduce their experimental setup and design, and then validate their countermeasure with Xilinx field programmable gate arrays which are configured with the pure AES circuit and the infected AES circuits. The power proportions (PPs) of HTs in the different infected AES circuits are various. The experimental results indicate that their proposed countermeasure can clearly detect HT with 0.14% small PP.

1 Introduction

Among the emerging security threats aiming at integrated circuits (ICs), hardware Trojan (HT) [1, 2] is an effective attack method because it is well stealthy and can implement various malicious behaviours [3–5]. Along with the globalisation of semiconductor industry, IC designers are forced to utilise third-party Internet protocol (IP) cores and outsource their designs to third-party foundries. Untrusted IP vendors can insert the so-called HT module into those IP cores and malicious foundries can insert the HT through lithography or doping modification. For the latter, academia has proposed several countermeasures in the past 11 years. This section first introduces these countermeasures including both advantages and disadvantages. Then, our motivation and paper organisation are presented.

1.1 Historical review

The proposed countermeasures in the past 11 years are trustworthy design, reverse engineering, functional test and side-channel analysis. Trustworthy design [6–9] is to insert special modules or circuits can prevent the HT insertion in the design phase or increase the HT detection rate during the detection phase. These modules or circuits can prevent the HT insertion in the manufacturing phase or increase the HT detection rate during the detection phase. However, the insertion of trustworthy modules or circuits may decrease the performance of the target chips because their optimum place and route are changed. Reverse engineering can absolutely confirm whether the target chips are infected but it is intrusive, which means the target chips will be destroyed after detection. In addition, its process is the most complicated, which means both the time cost and the economy cost are expensive. Functional test [10, 11], originating from automatic test pattern generation, is impossible to iterate every test vector because of the time limitation. How to generate the specialised test vectors for HT is its current bottleneck. Compared to other countermeasures, side-channel analysis [1, 12–16] has become the more popular and more effective one since 2007. Several physical signals such as power, electromagnetism, delay and temperature are widely used. However, the signal acquisition areas of the countermeasures using power and electromagnetism are too large so that the faint HT signals are prone to be merged in other signals. The countermeasure using delay needs to iterate every critical path to ensure whether the HT exists, which can lead to expensive time cost.

Thermal map is a typical temperature signal. It is more helpful because it can mitigate the influence of process variation (PV) [15]. Nowroz et al. [14] first proposed a countermeasure using thermal map as a side-channel signal. The simulation results indicate that the methodology using thermal maps in [14] can successfully detect the HT with 0.443 μW/μm² local Trojan power density (LTPD) under 30% PV in the AES benchmark, whose power density is 2.755 μW/μm². The simulation results also indicate that its another methodology using power maps can successfully detect the HT with 0.297 μW/μm² LTPD under 40% PV in the same Advanced Encryption Standard (AES) benchmark. However, the inversion from thermal maps to power maps needs the thermal resistance matrixes. Acquiring these matrixes of Application Specific Integrated Circuit (ASIC) need the expensive laser scanning system or the specialised simulation [17]. The latter estimation results will deviate from the real ASICs because the influence of PV cannot be coupled in the simulation. Gao et al. [15] proposed a two-level temperature difference framework based on thermal maps, which can detect the HT with 10⁻³ power proportion (PP) magnitude in real field programmable gate arrays (FPGAs). However, its experimental results indicate that the HT signal still can be partly merged after Kalman filtering. This phenomenon will lead to confusion in the practical application.

1.2 Motivation

In this paper, the methodology in [15] is further developed using principal component analysis (PCA)-based spatial projection transformation. Spatial projection transformation is a geometrical concept. It means constructing a novel coordinate system from the primitive coordinate system. In the novel coordinate system, some signal components, which are impossible to be recognised in the primitive one, can be easily distinguished. Although the thermal maps based countermeasures can mitigate PV’s impact, in the general time–space dimension, the faint HT signals still can be merged in the noises from environment, measurement and PV. Spatial projection transformation can contribute to better expose the HT from these noises in a more helpful coordinate system.
The organisation of this paper is as follows. In Section 2, we illustrate our methodology. A general HT detection framework based on difference temperature matrix is proposed to benefit the further development in the future. In addition, the PV mitigation mechanism of the countermeasure using thermal maps is presented. Section 3 indicates how PCA achieves spatial projection transformation and how it works to detect HT. In Section 4, we introduce our experimental setup, design and result. The result shows that our proposed approach can detect HT with 0.14% small PP in the AES circuit. Finally, Section 5 concludes our work and presents our future work.

2 Methodology

One may argue that whether the infected circuits can be differentiated from the others directly from the thermal maps of the target chips. The essence of thermal maps is temperature matrices. The value of these matrices' every cell means the measured temperature of the corresponding very small area in the target chips. Although HT implements malicious behaviours, its essence is circuits, as same as any and all normal modules in the target chips. When the target chips are working for detection, both HT and normal modules emit thermal signals present temperature value cells in the final temperature matrices. Hence, testers cannot make a judgement merely from these measured temperature matrices (i.e. measured thermal maps). Fig. 1 is a measured thermal map from an experiment. The $x$ and $y$ axes can help to determine the circuits' position coordinates. The colour temperature legends can help to determine the temperature values of different coordinates. The HT locates the area around $[180, 280]$ that cannot be directly found differences from the others. In this section, general detection framework using temperature matrices is first demonstrated, which is helpful for the further development in the future. Second, a deep discussion result about PV mitigation is presented.

2.1 Detection framework

Fig. 2 formulates the general HT detection framework based on difference temperature matrix. In this framework, both the thermal capture system and the simulation system are needed. The former consists of thermal camera and software, which measures the temperature matrices of the target chips. The latter includes NC Verilog, Primetime-PX, Hotspot etc., which generates the corresponding golden models for the target chips.

The value of a normal cell in a measured temperature matrix can be demonstrated by the following equation:

$$T_C = T_{C, \text{measurement}} + T_{C, \text{environment}} + T_{C, \text{circuits}} + T_{C, \text{process}} + eT_{C, \text{round}}$$

$T_C$ is the total temperature of the cell. $T_{C, \text{measurement}}$ and $T_{C, \text{environment}}$ are, respectively, the measurement noise and the environment noise from the measurement process. Both of them are Gaussian white noise. $T_{C, \text{circuits}}$ is the temperature caused by the operation of normal circuits with typical process parameters.
If the corresponding region of a cell in the target chip is infected by a Trojan, where the circuit to be tested and the environmental heat dissipation ability of the target chips, the value of the difference temperature value of every cell can be demonstrated by the following equation:

\[ T_{C} = T_{C\text{-measurement}} + T_{C\text{-environment}} + eT_{C\text{round}} \]  

(2)

If the corresponding region of a cell in the target chip is infected by HT, (1) should be changed to the equation below:

\[ T_{C} = T_{C\text{-measurement}} + T_{C\text{-environment}} + T_{C\text{circuits}} + T_{C\text{-process}} + eT_{C\text{round}} + T_{\text{ Trojan}} \]

(3)

where \( T_{\text{ Trojan}} \) is the temperature caused by HT.

The normal value of a cell in a golden temperature matrix can be demonstrated by the following equation:

\[ T_{\text{CGM}} = T_{\text{CGM\_circuits}} \]

(4)

where \( T_{\text{CGM\_circuits}} \) is the temperature caused by the operation of normal circuits with typical process parameters. If the corresponding region of a cell in the golden model does not have any circuits, the value of \( T_{\text{CGM}} \) is 0.

As Fig. 2 demonstrates, through the difference operation between the measured temperature matrices and the golden temperature matrices, the difference temperature matrix can be gained. The difference temperature matrix demonstrates the difference temperature value of every cell (\( C_{1}, C_{2}, \ldots, C_{mn} \)) at every time (\( t_{1}, t_{2}, \ldots, t_{p} \))

\[ \Delta T = \begin{bmatrix} \Delta T_{11} & \Delta T_{12} & \ldots & \Delta T_{1n} \\ \Delta T_{21} & \Delta T_{22} & \ldots & \Delta T_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta T_{m1} & \Delta T_{m2} & \ldots & \Delta T_{mn} \end{bmatrix} \]

The difference temperature value of a normal cell under one time can be demonstrated by the following equation:

\[ \Delta T_{C} = T_{C\text{-measurement}} + T_{C\text{-environment}} + T_{C\text{-process}} + eT_{C\text{round}} \]  

(5)

If the corresponding region of a cell in the target chips does not have any circuits, (5) can be simplified to the equation below:

\[ \Delta T_{C} = T_{C\text{-measurement}} + T_{C\text{-environment}} + eT_{C\text{round}} \]

(6)

If the corresponding region of a cell in the target chip is infected by HT, (5) should be changed to the equation below:

\[ \Delta T_{C} = T_{C\text{-measurement}} + T_{C\text{-environment}} + T_{C\text{circuits}} + T_{C\text{-process}} + eT_{C\text{round}} + T_{\text{ Trojan}} \]

(7)

Finally, detection algorithms (such as Kalman filtering, PCA, intelligent computing etc.) process the difference temperature matrices so that HT signals can be distinguished from the others. To reduce the influence of circuits around the cell’s corresponding region in the target chip, we should accelerate the heat dissipation during the measurement. However, this problem is not discussed in this paper.

2.2 PV mitigation mechanism

The practical has proven that PV can cause more significant effects than the HT, especially in the side-channel analysis using power and electromagnetism. Fig. 3 demonstrates the minimum signal acquisition area of countermeasures using different side-channel signals. The larger the acquisition area is, the more significant effects PV have than the HT. The countermeasure using electromagnetism uses electromagnetic probe to measure the target chips. The acquisition area of this probe is from \( 10^{-7} \) to \( 10^{-5} \) cm². Therefore, the impacts of PV using electromagnetism is the \( 10^{-5} \) of those using power. Similarly, decided by thermal cameras’ pixel size in the charge coupled device, the minimum signal acquisition scale using difference temperature matrix is micron, which means the impacts of PV is the \( 10^{-5} \) of those using power. Actually, in the target chips, a thermal cameras’ pixel size area only can include \(<20 \) gates. The exact gate number is decided by both the thermal camera and the chip technology. Hence, the countermeasure using difference temperature matrix is more effective on PV mitigation than the others.

3 HT detection using spatial projection transformation

Section 2 demonstrates our proposed detection framework, which indicates that detection algorithms are the kernel component. Our motivation to use spatial projection transformation has been presented in Section 1.2. This section first introduces how PCA can mathematically implement spatial projection transformation, and then reveals how PCA can be used in HT detection.

3.1 PCA-based spatial projection transformation

PCA [18] is a statistical approach. It can convert the correlated variables into linear uncorrelated variables through orthogonal transformation. Geometrically, this transformation can be treated as spatial projection transformation. Through this transformation, the mixed data that are impossible to be differentiated in the original coordinate system become easy to be distinguished under the new coordinate system. Although it is hard to define the physical meanings of every coordinate axis in the new coordinate system, the first several transformed features, namely the principal components, can reflect the information of original features as much as possible. Therefore, the useful dimension of transformed features can be decreased. Additionally, the independence of transformed features is better. The following demonstrates the spatial projection transformation principle of PCA.

Assume that there are \( n \) samples and each sample has \( p \) features. Here, matrix \( X \) demonstrates these \( n \) samples and their features

\[ X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = (X_{1}, X_{2}, \ldots, X_{p})^{T} \]  

(8)

Our purpose is to construct a new linear combination \((F_{1}, F_{2}, \ldots, F_{p})^{T}\) which can satisfy the following conditions:

IET Inf. Secur.
© The Institution of Engineering and Technology 2018
The quadratic sum of original features' coefficients is 1
\[ u_0^T + u_1^T + \cdots + u_p^T = 1 \]

(i) The new features are independent from each other
\[ \text{Cov}(F_i, F_j) = 0; \quad i \neq j; \quad i, j = 1, 2, \ldots, p \]

(ii) The variances of new features decrease in sequence
\[ \text{Var}(F_1) \geq \text{Var}(F_2) \geq \cdots \geq \text{Var}(F_p) \]

The bigger the variance is, the more principal the component \( F_i \) is.

It is easy to find that the critical problem of PCA is to calculate the coefficients
\[ u_i = (u_{i0}, u_{i1}, \ldots, u_{ip})^T; \quad i = 1, 2, \ldots, p \]

and then implement spatial projection transformation.

### 3.2 PCA-based HT detection

On the basis of the introduction of Section 2.1, a difference temperature matrix is calculated from an experiment. For better understanding, this matrix is visually demonstrated as Fig. 4. Its x-axis is the time (frame) and its y-axis is the difference temperature (°C). This figure records every cell's difference temperature variation along with time, so the space perpendicular to the x-axis stands for the space dimension of the target chip. It is easy to find the fact that the faint HT signal is merged in the noises, which indicates that the HT signal cannot be exactly distinguished in the general time–space dimension. Hence, a novel coordinate system needs to be constructed, which can expose HT.

### 4. Experimental setup, design and result analysis

In this section, we validate the effectiveness of our proposed approach. The experimental setup and design are separately introduced, from which the results are gained and analysed.

#### 4.1 Experimental setup

In our experimental setup, 7 Xilinx Spartan-3A XC3S50 FPGAs are used to evaluate our approach. This type of FPGA adopts ball grid array packaging that is one type of flip-chip packaging. Its configurable logic blocks (CLBs) contain flexible look-up tables and can perform a wide variety of logical functions [5]. The AES benchmark can be implemented with the CLBs and constrained by PlanAhead in a fixed layout. Finally, we use incremental compilation to add the HT to form a new layout without changing previous fixed layout. In our experiment, the FPGAs' package heat spreaders are removed, and an forward looking infrared camera is utilised to capture thermal patterns, with 25 × 25 μm² spatial resolution, 25 Hz operation frequency and 30 mK noise equivalent temperature difference. To capture clear thermal maps, a cooling fan is utilised to expedite detected FPGAs' heat dissipation. Fig. 5 demonstrates our experimental setup.

#### 4.2 Experimental design

In the seven FPGAs, one FPGA is configured with the pure AES circuit to mimic the golden model. Three FPGAs (FPGA1, FPGA2, FPGA3) are configured with the infected AES circuits to mimic the ASIC inserted with HTs by third-party foundries. The other three FPGAs (FPGA4, FPGA5, FPGA6) are configured with the pure AES circuit to mimic the pure chips after the third-party manufacturing. The power of the pure AES circuit in the FPGAs is 0.765 W. The HT in our experiment is shown in Fig. 6. This HT which can implement information leakage consists of trigger logic (TL) and payload logic (PL) (eight gates). The TL monitors target signal and triggers PL as soon as the specific signal is appearing.

Almost all the power of the HT is consumed by its TL because the PL works only when the specific signal appears. Therefore, the
power of the HT can be controlled through adjusting its monitored signals frequency. The HTs’ powers of different infected AES circuits are presented in Table 1. Due to the fact that the temperature caused by the operation of circuits is proportional directly to the power of the circuits, we use the PP of HT to evaluate the detection ability of our proposed approach, and the HTs’ power of different infected AES circuits are various (as shown in Table 1).

4.3 Result analysis

According to the theory proposed before, the difference temperature matrixes of the six tested FPGAs can be gained, and then PCA can be used to construct novel coordinate systems for them for exposing the HT. Table 2 shows the contribution rates of the first three components of the six difference temperature matrixes, from which it is easy to find that the sums of every matrix’s first three component contribution rates are bigger than 80%. Therefore, the first three components in our experiment can reflect excellently the nature of the original matrixes. The novel coordinate systems constructed by the first three components and the spatial projection transformation results of six tested FPGAs are illustrated in Fig. 7, whose x-axis, y-axis and z-axis stand for the first component, the second component and the third component, respectively.

Although the transformed spatial projections of normal circuits have some outlier points in the results of FPGAs configured with the infected AES circuits, HTs in all of the three FPGAs are detected exactly (as shown in the red circles of Figs. 7a–c). For the results of FPGAs configured with the pure AES circuit, their transformed spatial projection does not appear HT outlier point (as shown in Figs. 7d–f). Therefore, the experiment results indicate our proposed approach can exactly detect HT with 0.14% small PP.

### Table 1 Powers and PPs of HTs in different infected AES circuits

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Power, μW</th>
<th>PP, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPGA1</td>
<td>153.00</td>
<td>0.20</td>
</tr>
<tr>
<td>FPGA2</td>
<td>191.25</td>
<td>0.25</td>
</tr>
<tr>
<td>FPGA3</td>
<td>107.10</td>
<td>0.14</td>
</tr>
</tbody>
</table>

### Table 2 First three components’ contribution rate of every tested FPGA

<table>
<thead>
<tr>
<th>FPGA</th>
<th>Configured with infected AES circuits</th>
<th>Configured with the pure AES circuit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st component</td>
<td>2nd component</td>
</tr>
<tr>
<td>FPGA1</td>
<td>99.3598</td>
<td>0.0291</td>
</tr>
<tr>
<td>FPGA2</td>
<td>87.6665</td>
<td>1.3153</td>
</tr>
<tr>
<td>FPGA3</td>
<td>99.8772</td>
<td>0.0073</td>
</tr>
<tr>
<td>FPGA4</td>
<td>95.4919</td>
<td>0.2145</td>
</tr>
<tr>
<td>FPGA5</td>
<td>99.3731</td>
<td>0.0331</td>
</tr>
<tr>
<td>FPGA6</td>
<td>87.1660</td>
<td>0.5766</td>
</tr>
</tbody>
</table>

![Fig. 7 Experimental results](image.png)

(a) Spatial projection transformation result of FPGA1, (b) Spatial projection transformation result of FPGA2, (c) Spatial projection transformation result of FPGA3, (d) Spatial projection transformation result of FPGA4, (e) Spatial projection transformation result of FPGA5, (f) Spatial projection transformation result of FPGA6
5 Conclusion and future work

In this paper, we first propose our methodology which introduces the general HT detection framework based on difference temperature matrix and the PV mitigation mechanism. Then, we present to detect HT using spatial projection transformation, in which how PCA can implement spatial projection transformation and detect HT is demonstrated. Finally, the experimental setup and design are presented. The experimental results show that our proposed approach can successfully detect HT with 0.14% small PP in AES circuits. Compared to [15], the experiment in this paper is more practical because it adopts veritable cryptographic circuits (AES) and a real HT with eight gates, which can implement information leakage. In addition, the experimental results indicate that all the HTs are clearly detected without obfuscation.

For our future work, we will try to use liquid nitrogen to accelerate the heat dissipation of target chips, and further validate our countermeasure in real ASICs with the pure AES circuit and the infected AES circuits.

6 Acknowledgment

The author thank the support of Major State Basic Research Development Program of China (No. 61331604).

7 References