ELectronics EXpress

LETTER

IEICE Electronics Express, Vol.16, No.2, 1–8

# Golden chip free Trojan detection leveraging electromagnetic side channel fingerprinting

**Jiaji He**[1a)], **Yanjiang Liu**[1], **Yidong Yuan**[1,2], **Kai Hu**[3], **Xianzhao Xia**[1], **and Yiqiang Zhao**[1]

[1] *School of Microelectronics, Tianjin University, Tianjin 300072, China*

[2] *Beijing Smartchip Microelectronics Technology Company Limited, Beijing 100192, China*

[3] *School of Electrical and Electronic Engineering, Tianjin University of Technology, Tianjin 300384, China*

a) *dochejj@tju.edu.cn*

**Abstract:** Side channel based hardware Trojan detection is one of the most investigated schemes to ensure the trustworthiness of integrated circuits (ICs). However, nearly all of the side-channel methods require a golden chip reference, either a trusted fabricated circuit or layout, which is very difficult to access in reality. In this paper, we propose a golden chip free electromagnetic (EM) side channel simulation and statistical Trojan detection framework. We utilize the design data at early stage of the IC lifecycle to generate EM radiation, and the generated EM traces serve as the golden reference. In order to leverage the EM signatures, a neural network algorithm is utilized for Trojan detection. Experimental results on selected AES benchmarks on FPGA show that the proposed method can effectively detect Trojans with the presence of noise and variations.

**Keywords:** hardware Trojan detection, golden chip free, electromagnetic side channel, neural network algorithm

**Classification:** Integrated circuits

## References

[1] M. Tehranipoor and F. Koushanfar: "A survey of hardware Trojan taxonomy and detection," IEEE Des. Test Comput. **27** (2010) 10 (DOI: 10.1109/MDT.2010.7).

[2] M. Oya, *et al.*: "A hardware-Trojans identifying method based on Trojan net scoring at gate-level netlists," IEICE Trans. Fundamentals **E98A** (2015) 2537 (DOI: 10.1587/transfun.E98.A.2537).

[3] M. Oya, *et al.*: "Hardware-Trojans rank: Quantitative evaluation of security threats at gate-level netlists by pattern matching," IEICE Trans. Fundamentals **E99A** (2016) 2335 (DOI: 10.1587/transfun.E99.A.2335).

[4] X. Xie, *et al.*: "Hardware Trojans classification based on controllability and observability in gate-level netlist," IEICE Electron. Express **14** (2017) 20170682 (DOI: 10.1587/elex.14.20170682).



DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

[5] C. Wang, *et al.*: "An intelligent classification method for Trojan detection based on side-channel analysis," IEICE Electron. Express **10** (2013) 20130602 (DOI: 10.1587/elex.10.20130602).
[6] J. He, *et al.*: "Hardware Trojan detection through chip-free electromagnetic side-channel statistical analysis," IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **25** (2017) 2939 (DOI: 10.1109/TVLSI.2017.2727985).
[7] D. Agrawal, *et al.*: "Trojan detection using IC fingerprinting," Proc. IEEE Symposium on Security and Privacy (SP) (2008) 296 (DOI: 10.1109/SP.2007.36).
[8] B. Zhou, *et al.*: "Cost-efficient acceleration of hardware Trojan detection through fan-out cone analysis and weighted random pattern technique," IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **35** (2016) 792 (DOI: 10.1109/TCAD.2015.2460551).
[9] F. Menichelli, *et al.*: "High-level side-channel attack modeling and simulation for security-critical systems on chips," IEEE Trans. Depend. Secure Comput. **5** (2008) 164 (DOI: 10.1109/TDSC.2007.70234).
[10] L. K. Hansen and P. Salamon: "Neural network ensembles," IEEE Trans. Pattern Anal. Mach. Intell. **12** (1990) 993 (DOI: 10.1109/34.58871).
[11] Trust-HUB. https://www.trust-hub.org/ [online; accessed Dec-12-2018].

## 1 Introduction

Hardware Trojans (HTs) are malicious hardware modifications to ASICs, commercial-off-the-shelf parts, micro-processors, micro-controllers, network processors, digital-signal processors or IoTs. Most of the Trojans can be digitally triggered and can be further divided into combinational and sequential types [1]. Typically, sequential Trojans are usually related with main clocks or finite state machines of the original circuit. To compromise the threats introduced by HTs, HT detection approaches have been proposed at pre- and post-silicon stages of the whole IC lifecycle. In the pre-silicon stage, the gate-level netlists are analyzed to find abnormal or extra Trojan nets [2, 3, 4]. These pre-silicon Trojan detection approaches provide a reliable guarantee for the trustworthiness of the circuits before tape out. In the post-silicon stage, however, the side channel based methods [5, 6, 7] are more flexible and easy to implement, where I/O ports and side-channel parameters are utilized to find abnormal behaviors introduced by the Trojans. While various HT detection approaches have been explored by many researchers, statistical side-channel analysis has been among the most heavily investigated ones [8]. However, most side-channel methods rely heavily on the existence of a trusted golden chip or other profiles alike, such as golden layout. Absence of a reliable fabricated golden chip or golden layout makes practical applications of side-channel detection approaches unfeasible.

In this paper, we propose a golden chip free electromagnetic side-channel simulation and statistical Trojan detection framework. In the EM modeling and simulation process, only the genuine RTL design is required for generating the circuit's EM radiation. The simulated EM traces are transformed into frequency domain to leverage the spectral features of the EM radiation exempt from the influence of noise and variations. Finally, a pattern recognition shallow neural


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

network is utilized to extract and learn the EM signatures, which are features of the EM spectra for Trojan detection. The main contributions of this paper are as follows.

- A trusted EM model is established utilizing the RTL design data, and the generated EM traces serve as the golden reference in Trojan detection;
- Optimization of the EM model is made towards physical implementations by considering the actual executions of the circuit, and the simulated EM model matches with actual measurements well;
- EM spectral features are extracted and learned by a neural network for Trojan detection.

## 2 Background

We assume that the RTL design data is trusted, or if the design data is not trusted, the trusted circuit's functional simulation data is available, either without Trojans or with Trojans dormant. We also assume the circuits have certain clock signals, because the inner logic changes are key for generating the EM traces. We primarily focus on the HT detection of sequential Trojans in this paper, because the method achieves the best results due to the Trojans' relations with the original circuit's clock and inner logic values.

### 2.1 HTs and EM radiation

EM radiation arises as a consequence of current flows within control, I/O, data processing or other parts inside a chip. In real chips, current only flows when there are changes in logic states, thus the EM radiation carries information about the currents and hence the events and relevant states inside the chips. Clearly, HTs are modifications to original circuits and HTs usually consist of trigger parts and payload parts. The trigger parts typically have strong relations with clock signals, FSMs or state nodes in the original circuits, thus will generate strong EM radiation. The payload parts are responsible for conducting malicious functions. Also, when the payload parts take effect, they will generate strong EM radiation. Even if the HTs' trigger parts remain silent, they will still monitor the internal signals and influence the current flows within the IC, thus they will also affect the EM radiation of the IC. Further, the structural changes in the IC, which are introduced by HTs, will cause the variations in leakage currents, which will also alter the EM radiation. With the help of state-of-the-art data processing techniques, we neither require the Trojan to be brought to the triggering state, nor the effect of the Trojan payload to be observed.

## 3 Golden chip free EM spectrum modeling and Trojan detection methodology

In this section, we discuss the overall framework, algorithms and steps included in the EM radiation modeling and Trojan detection methodology. The overall framework is demonstrated in Fig. 1. The whole framework has three steps. In the first step, the genuine RTL design and known Trojan-infected RTL design are feed into the EM simulation model to obtain the training spectra for the neural network. In the second step, EM radiation is collected from the fabricated chips under test and


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

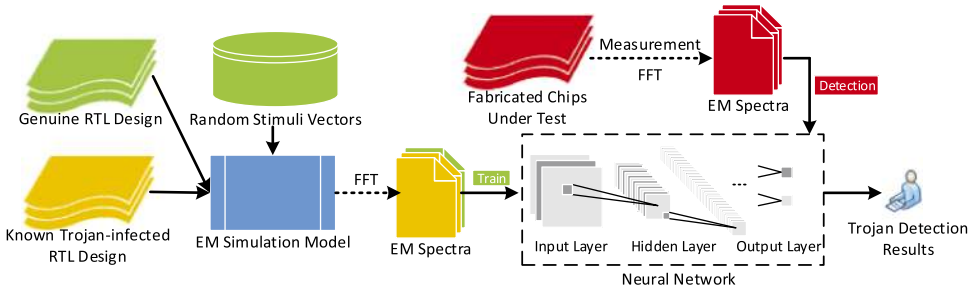


**Fig. 1.** Golden-chip free Trojan detection framework

is transformed to frequency domain. In the third step, the trained neural network is utilized to detect whether the chips under test are Trojan infected or not.

### 3.1 RTL EM side channel radiation model

Concerning the simulation of an IC's EM side-channel radiation, a few papers have put forward some ideas using Hamming Distance, Hamming Weight or improved models [9]. The existing models have several drawbacks, including not scalable for very large designs, not considering real hardware implementations, etc. The model in this paper is built based on Hamming Distance, and the simulated traces are modeled with specific factors of circuits. The factors that contribute most to the EM radiation, such as data transitions and driving capability, are taken into consideration, while factors that have little impacts for direct near-field EM radiation, like coupling effect, are ignored. Through optimum seeking of factors, major parts of the radiation which caused by signal transitions are modeled.

The model is designed for matching with actual test data. The main contributors for EM radiation will be data transitions and driving capabilities. Specific optimization is made targeting FPGA implementation by taking the actual executions of the circuit into consideration. In the FPGA implementation, the initial and final states of the $j$-th register/LUT are denoted as $M_j$ and $N_j$ respectively, and $t$ represents the moment of the transition. A script written in *tcl* language is utilized to calculate registers, LUTs and driving capabilities. The fan-out number of the $j$-th register or LUT is denoted as $F_j$, then the simulated EM side-channel trace $R(t)$ is modeled as Equation (1), where $\oplus$ denotes the exclusive OR operation.

$$R(t) = \sum_{j=1}^{n} F_j \times (M_j \oplus N_j) \tag{1}$$

All results from Equation (1) are added up along the time axis to get the simulated traces in time domain with every fan-out number as their weights. Also if the stimuli changes, the simulated traces vary accordingly. The simulated EM signal can serve as the "golden-reference" for EM side-channel based Trojan detection. With the trusted RTL source code, we apply the source code in our EM model to generate the golden EM data. Then Fast Fourier Transform is applied on the EM data to get the EM spectrum.

### 3.2 Golden chip and HT EM radiation construction

The principal basis of the golden chip free EM side-channel based HT detection methodology is to find the differences between the simulated trace and the


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

measured traces from chips under test. The signal in time domain is $R(t)$, and its corresponding expression in frequency domain is denoted as $S(\omega)$. According to Fourier transform we have Equation (2), where $\omega$ is its corresponding frequency.

$$S(\omega) = \mathcal{F}(R(t)) = \int_{-\infty}^{+\infty} R(t)e^{-j\omega t}dt \tag{2}$$

The detailed composition of $\mathcal{F}(R(t))$ signal captured by the probe includes main clock and its harmonics, whose frequency can be denoted as $g_1$, $g_2 \cdots g_g$ respectively, some periodic signals generated by the circuits, whose frequency can be denoted as $f_1$, $f_2 \cdots f_f$ respectively, and other unintended signals, including process variations and noises, denoted as $U$. Assuming a sequential HT with signal transition frequency $T_1$ is inserted into the circuit, under the same circumstances and after FFT, the EM signals captured by the probe are formulated as Equation (3), where $A_{1i}$, $A_{2i}$, $A_3$ and $A_4$ denote the magnitude of each frequency components, respectively. Considering the Trojan's influence $A_4S(jT_1)$, we are able to separate different components inside the circuit. We can utilize the features of the Trojan's radiation components, either trigger parts or payload parts, to detect Trojans from golden EM model.

$$\begin{aligned}\mathcal{F}(R(t)) = &\sum_{i=1}^{g} A_{1i}S(jg_i) + \sum_{i=1}^{f} A_{2i}S(jf_i) \\ &+ A_3S(jU) + \mathrm{A}_4S(jT_1)\end{aligned} \tag{3}$$

### 3.3 Data processing and neural network based Trojan detection

The measured EM radiation is exposed to all kinds of noises, so denoising process is needed to optimize the data. When measuring traces through experiments, the traces are averaged using the oscilloscope to eliminate most of the random noise. After the data is acquired using the oscilloscope, further denoising is performed to reduce noise and rise signal-to-noise ratio (SNR) [6]. After denoising step, the EM data is transformed into frequency domain using FFT. To fully utilize the abundant spectral features of the EM spectra, the neural network is the most suitable algorithm for retaining and extracting EM radiation signatures [10]. Neural network's application in the field of HT detection is still in exploratory stage, but it has a strong ability of nonlinear mapping and adaptive learning ability. To be more comprehensive, other than the simulated genuine circuit, a few representative known Trojan-infected benchmarks are also simulated in the EM model to train the neural network.

The EM model's output $\mathcal{F}(R(t))$ is used as the inputs for the neural network. More specifically, the FFT of $R(t)$ matrix, $S(w)$, is the input signal for the neural network. In the training process of the neural network, the size of the input layer is $N$, which is the dimension of $S(w)$. The hidden layer has $M$ neurons and the output of the $j$-th neuron of the hidden layer is denoted as $H_j$. The output signal $O_k$ is described as Equation (4), where $W_{ij}$ is the weight between input layer and hidden layer and $Z_{jk}$ is the weight between hidden layer and output layer. The topology of the neural network is illustrated in Fig. 2. Through utilizing existing pattern


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

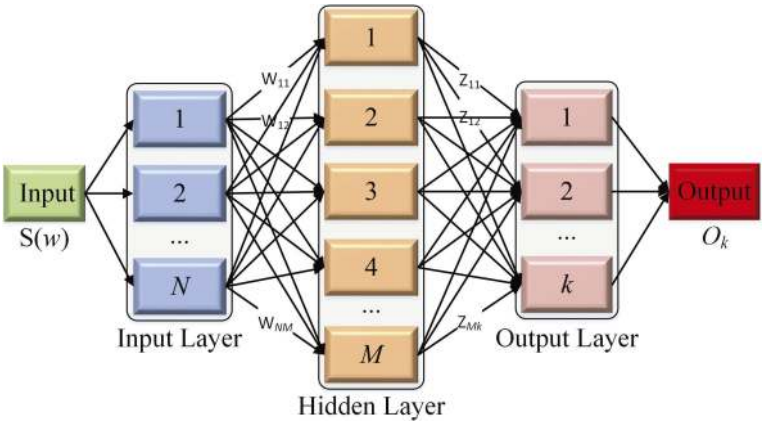


**Fig. 2.** Neural network topology

recognition tools[1], we train the network to update the weights until it achieves convergence. After the neural network achieves convergence, it is utilized for Trojan detection, and then the data collected from chips under test are applied as inputs to detect Trojans by checking the outputs of the trained neural network.

$$O_k = \sum_{j=1}^{M} H_j Z_{jk} = \sum_{j=1}^{M} \left\{ \sum_{i=1}^{N} W_{ij} S(w)_i \right\} Z_{jk} \tag{4}$$

## 4 Experimentation and Trojan detection

To validate the proposed EM radiation modeling methodology and Trojan detection framework, we carried out real hardware Trojan detection experiments on a FPGA platform. Several data processing steps are carried out to find the differences in simulated spectrum and actual spectra as discussed in Section 3.3.

### 4.1 Experiment setup

The experiment setup is shown in Fig. 3. The experiment platform is a SAKURA-G FPGA board specifically designed for research and development on hardware security. Two Spartan FPGAs are integrated on the board. The input operands are provided by the controller FPGA to the main FPGA. The main FPGA is in charge of conducting out operations and will not be affected by other parts on the board. A LANGER EM near field probe, which is fixed on the board, is utilized to acquire EM radiation. After acquiring EM radiation by the near field probe, the signals are amplified using a pre-amplifier PA303 up to 30 dB magnification. Then the signals are collected and transferred to the computer for further analysis.

### 4.2 Trojan detection results

Explicitly, we choose 5 categories of AES circuits downloaded from the Trust-HUB online repository [11] as benchmarks. The Class 1 (genuine AES) represents the original circuit, Class 2 (AES-T100 & T200) represents data-leak Trojans through capacitance, Class 3 (AES-T1600 & T1700) represents data-leak Trojans through antenna, Class 4 (AES-T1800 & T1900) represents denial-of-service type

[1]We only utilize the neural network algorithm for Trojan detection, however, how to design and optimize the neural network is out of the scope of this paper.


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

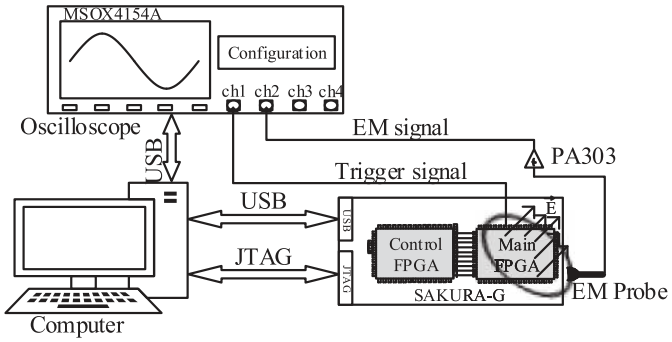


**Fig. 3.** Experiment setup

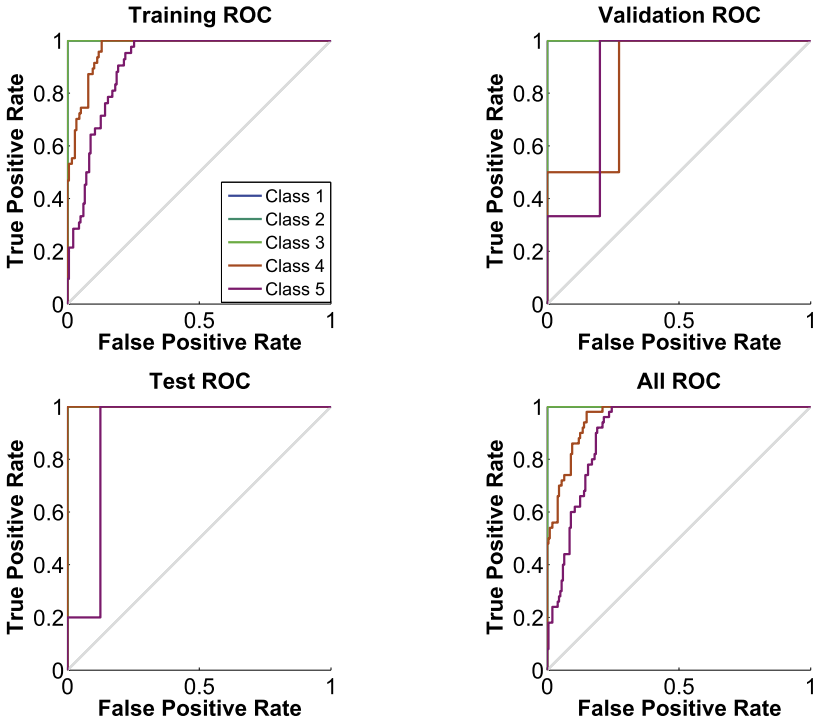


**Fig. 4.** Neural network training results

Trojans, and Class 5 (AES-T2000 & 2100) represents combinational and sequential data-leak Trojans. In the training process, we randomly pick 1000 simulated genuine AES traces, which are used as the trusted reference bundle. Also, 1000 simulated traces each from AES-T100, AES-T1600, AES-T1800 and AES-T2000 are picked as known Trojan-infected benchmarks. For the purpose of evaluating the EM model and Trojan detection, the output of the neural network is set as five classes, which are genuine and different Trojan-infected classes. A pattern recognition shallow neural network toolkit in MATLAB is utilized through the *nprtool* GUI. The number of hidden layer is set as one, then an iteration script is developed to change the neurons in the hidden layer to achieve the best training result. More specifically, 70% samples are used for training, 15% samples are used in validation, and 15% samples are used during testing to measure the network performance. The neural network is trained following a scaled conjugate gradient method, and the best validation performance is reached at epoch 24. The training results are shown in Fig. 4. The neural network has three layers and ten hidden neurons.


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019

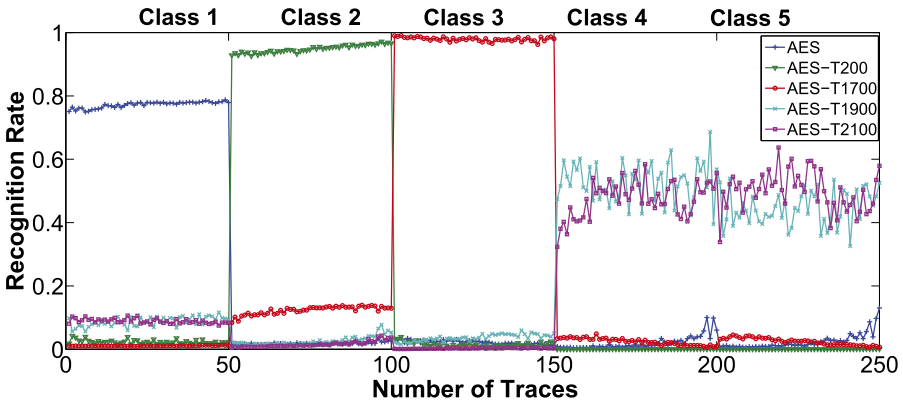


**Fig. 5.** Trojan detection results

The detection results are shown in Fig. 5. From the detection results of Class 1, the original AES circuit's recognition rate is over 80%, while other circuits' contributions in the Class 1 are significantly lower. This observation confirms two conclusions: the first is that our EM simulation model matches with real measurements well with the presence of noise and variations, and the second is that our proposed EM detection framework can clearly distinguish between genuine and Trojan-infected circuits. From the results of Class 2 and Class 3, the AES-T200 and AES-T1700 are totally separated from other HTs within each category. From the results of Class 4 and Class 5, although the AES-T1900 and AES-T2100 are mixed up by the tool, they are still distinguishable from other HTs. Overall, our framework can detect all Trojans, further, as a proof-of-concept, our framework can distinguish even different types of Trojans with an averaged 89.2% accuracy rate. Compared with the work in [6], there are fewer constraints on the Trojans and input vectors, besides, the proposed Trojan detection method is more applicable to real applications as the Trojans can be detected before activation.

## 5 Conclusion and future work

In this paper, we propose a hardware Trojan detection methodology using EM side-channel based spectrum modeling and statistical data analyzing. We demonstrate that the simulated EM spectrum can be used as a golden reference for HT detection, and the experimental results validate the effectiveness of our method. The framework in this paper makes side-channel based hardware Trojan detection more applicable and practical for real implementations.

## Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No. 61832018.


DOI: 10.1587/elex.16.20181065
Received November 30, 2018
Accepted December 12, 2018
Publicized December 27, 2018
Copyedited January 25, 2019