Exploiting Phase Information in Thermal Scans for Stealthy Trojan Detection

Maxime Cozzi, Jean-Marc J.-M. Galliere, Philippe Maurine

Exploiting Phase Information in Thermal Scans for Stealthy Trojan Detection

Maxime COZZI
University of Montpellier - CNRS
LIRMM
Montpellier, France
maxime.cozzi@lirmm.fr

Jean-Marc GALLIERE
University of Montpellier - CNRS
LIRMM
Montpellier, France
jean-marc.galliere@lirmm.fr

Philippe MAURINE
University of Montpellier - CNRS
LIRMM
Montpellier, France
philippe.maurine@lirmm.fr

***Abstract*—Infrared thermography has been recognized for its ability to investigate integrated circuits in a non destructive way. Coupled to lock-in correlation it has proven efficient in detecting thermal hot spots. Most of the state of the Art measurement systems are based on amplitude analysis. In this paper we propose to investigate weak thermal hot spots using the phase of infrared signals. We demonstrate that phase analysis is a formidable alternative to amplitude to detect small heat signatures. Finally, we apply our measurement platform and its detection method to the identification of stealthy hardware Trojans.**

***Index Terms*—Trojan Detection, Lock-in Thermography, Thermal Mapping, Phase**

I. Introduction

Over the years, many methods of investigation have been developed to assess the Hardware Trojan (HT) issue. The most efficient method in order to retrieve precisely the structure of an IC is mechanical reverse engineering. This implies mechanically and/or chemically etching the die to perform imaging reconstruction and analysis using a scanning electron microscope. This, of course, leads to complete destruction of the circuit and therefore corresponds to what is called a "destructive" method.

"Non destructive" techniques were imagined in order to implement on-line verification. They can be split into two branches: invasive, that requires modification of the circuit's layout to detect or prevent HTs [1], [2], and non invasive, that use features of the circuit to apply run-time, logic testing or side channel [3], [4]. We found the latter one particularly interesting as these methods aim at retrieving small changes in the physical features of the IC rather than attempting to trigger it. A more detailed review of existing HT detection methodology can be found in [5].

Side channel imaging consists in mapping a circuit's activity regarding a specific physical variation as a secondary effect of the die's operation. Common vectors of analysis are power consumption, electromagnetic emissions, etc. Because thermal dissipation is a direct consequence of any activity of an IC we focus, in this paper, on low power thermal investigation. Most of the work presented in the state of the Art focuses on amplitude thermal maps to retrieve activity areas on the die. This paper adopts a different approach, as we present a platform that uses the phase of the thermal measured signal in order to locate weak thermal activities. The idea is to take advantage of the great stability of circuits operation in time. This is then applied to detecting small stealthy HTs.

The organization of this paper is as follows. First we give a brief reminder of lock-in correlation and how it is applied to our IR platform. Then we present our platform and the implementation of the characterization and validation test-chip circuit. Finally, we apply our detection methodology in order to locate HTs mapped in our test-chip.

II. Lock-In Thermography

Because lock-in thermography is based on synchronous detection, heat induced by operating the IC must be modulated at a known frequency in order to generate an alternative heat pattern. This modulation, as restrictive as it may seem, can be implemented in numerous ways. The idea is to create a variation in the power consumption to highlight a specific area of the chip. For example, in the case of a micro-controller, the modulation can be implemented as a software load variation by alternating idle and processing sequences of a peripheral (memory write operation, arithmetic operation, etc). Another method widely applied in the state of the Art, uses the power supply to superimpose an AC component on top of the DC bias current. However, the latter method is often compromised in modern complex circuits by the presence of an internal regulator able to reject the power supply ripples. In our case, we assume the fundamental component of the heat modulation is expressed as follows in (1).

$$S_{fund} = A.sin(2.\pi.f_{lockin} + \Phi) \tag{1}$$

In order to retrieve the targeted signal, the output of the thermal sensor is computed in two different and independent channels:
- the first one multiplies and integrates the sensor's output by a sinus in phase with S_{fund}. The resulting signal is called S_0.
- the second one multiplies and integrates the sensor's output by a cosinus (90 ° phase shift). The resulting signal is called S_{90}.

From these two channels we can then retrieve the amplitude A and the phase Φ of the thermal signal using (2) and (3):

$$A = \sqrt{S_0^2 + S_{90}^2} \tag{2}$$

$$\Phi = Arctan(\frac{S_0}{S_{90}}) \quad (3)$$

At the end of this procedure, one has at disposal the amplitude of the signal which is prone to emissivity contrast and temperature drift errors, especially when aiming at detecting extremely low power heat sources. One also has access to the phase of the heat source; phase values which only depends on the IC operation, known to be extremely precise and stable in time especially when considering the time constants characterizing thermal behavior and those characterizing IC operations. Thus phase analysis seems to be the most robust information for low power heat sources detection.

III. Experimentation and Implementation

In this section, we describe the experiment considered in the rest of this paper, the hardware used to acquire thermal maps and how we emulated the insertion of a malicious circuit in an IC.

A. Thermal Acquisition

To detect thermal emissions, we use a InAs photodiode cooled at -60 ° C that can detect signals of wavelength between 1 and 3.8 μm. The output variable of the sensor being a photocurrent, we added a transimpedance amplifier with a gain of 2.10^8 V.A. A generic remote controlled oscilloscope is then used to digitalize the measured signal at $10\,kS/s$. While the sensor can theoretically detect inputs signals up to $1\,MHz$, the overall bandwidth is limited by the amplifier to $20\,kHz$.

B. Hardware setup

The targeted IC is a Xilinx Virtex 5 FPGA, designed in 65 nm CMOS technology and with direct access to the substrate. This FPGA is programmed to integrate a 128 bit AES hardware encryption circuit running at 200 MHz (system clock). The latter is used to generate background heat activity on the circuit and emulate "normal" operation of a die. To avoid any modification in the layout while reprogramming the chip, it was implemented as a hardware macro. This ensures that all slices are placed at a unique location when programming the FPGA even if changes are applied to the rest of the IC. In order to apply the lock-in correlation we create a heat modulation by toggling the clock signal of the AES. Because the following work is exploiting phase information contained in thermal signal we wished to minimize any source of jitter. In this optic, we chose to implement the lock-in frequency directly from the chip's clock signal instead of using an external signal generator. For that, a simple counter was used to create the f_{lockin} signal at the required frequency. In addition this method is mere representative of what can be done using a software modulation.

C. Trojan Emulation

According to the work presented in [7], a Trojan can be seen as two separate pieces of hardware:
- the trigger which is the part of the circuit waiting for the proper conditions to be met in order to activate the malicious function of the Trojan.
- the payload which is the circuit realizing the harmfull function of the Trojan when the specific conditions are met. Let us imagine an attacker counting AES ciphering activations and deciding to force the encryption key to a predetermined value after a chosen number of occurrences. In this basic example, the counter is the trigger as it is the part of the circuit that awaits for the specific conditions to be met. The payload is the piece of hardware forcing the use of a known key as it realizes the 'harmfull' part of the malicious circuit. In this paper, we are interested in detecting the trigger's circuitry. Because this part of the Trojan is waiting for specific conditions to be met, its power consumption should be constant. In the case of a synchronous Trojan it is reasonable to assume that it is clocked by the system's clock signal. As a consequence we chose to gate the system clock with the f_{lockin} signal to realize the heat modulation.

The Trojan himself was emulated by placing several microheaters together to generate a constant consumption. Those microheaters are Ring Oscillators (RO) constituted of three gates: two "not" gates and one "nand" gate. The nand gate allows us to generate an "enable" signal in order toggle the microheaters at the f_{lockin} frequency and thus emulating the heat generated by the power consumption of the trigger. Each RO has a power consumption of 250 μW and occupies one slice in the FPGA.

The case presented in the next section is an attacker trying to blend the Trojan's heat signature in the background activity of the circuit. In this case we say the Trojan is "stealthy". Fig. 1 presents the layouts of the FPGA in both cases.

IV. Trojan Detection

In this section, we aim at detecting a Trojan whose heat diffusion is blended in the one of another part of the circuit. For that, we apply the detection methodolodgy developped in [8] for amplitude, to phase measurements. This technique uses statistical tools in order to detect small variations in amplitude between a reference thermal map and the DUT measurements

A. Methodology

The method we use to detect HT is based on comparing phase values between a reference thermal map (a.k.a as golden

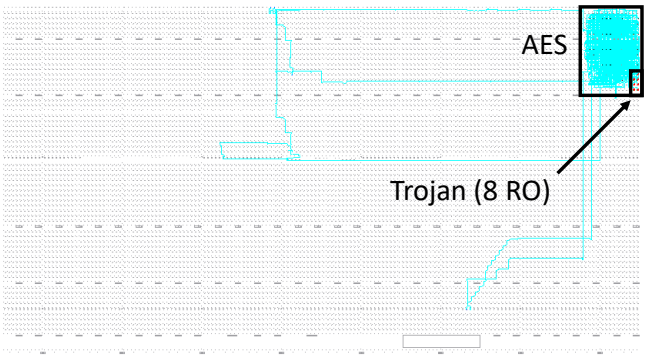

Fig. 1. FPGA layout for a stealthy Trojan insertion

thermal map) and the suspect circuit. Applying Welch's T-test to amplitude thermal maps has proven efficient in detecting very small variations in their statistical distribution [8]. In this paper we demonstrate that this technique is even more efficient when applied to phase values.

Before applying any statistical test, the raw data from the sensor needs to be freed of any outliers. Our data is composed of a hundred lock-in phase (respectively amplitude) acquisitions for a phase map of 80×80 pixels. To remove outliers, we compute the average $\bar{\Phi}$ and the standard deviation σ of the 100 measured phase values for each location (x,y) on the die. Any phase value exceeding $\bar{\Phi} + 3 \times \sigma$ is replaced by the median values of the 100 measures.

The next step consists of ensuring both circuits possess the same reference values. In other word, we make sure that external parameters drift, such as temperature variation or noise, cause minimal false positive when applying Welch's T-test. This is done by classical standardization of statistical data at the scale of the whole cartography, as described in [8].

Once this process is complete, the data are ready to be applied to Welch's T-test. Thermal measurements, acquired using the platform described in section III, provide two sets of n samples at each location (x,y) on the die. The first set of samples belongs to the chip that is being investigated while the second one corresponds to a trusted die and will serve as a golden reference. Therefore Welch's T-test can be applied between each pair of sample sets to detect variation between the golden statistical distribution and the DUT's one. The output variable of the T-test is the T-value and represents how different the two sets of measurements are. The higher (in absolute) the more difference there is between the two sample sets. Final results of the Trojan investigation are presented as pseudo-3D maps. The x and y direction corresponds to spatial coordinates of the DUT while the T-value is plotted over the z direction.

B. Highlighting Trojan Activity

Fig. 2 presents results of Welch's T-test. In this case, only the top half of the circuit has been mapped as the rest of the circuit presents no interest. The resolution of the presented images are 80×40 pixels, with $1\,pixel = 200 \times 200\,\mu m^2$. A_{ref} and Φ_{ref} show results of the T-tests between two golden thermal maps respectively for amplitude and phase. A_{HT} and Φ_{HT} present the results of the T-tests between a Trojan infected chip and a golden one. In this case, the HT is emulated by 4 ROs for a total power consumption of 1 mW. The 4 ROs are spread on 4 slices over a total of 17280 which represent

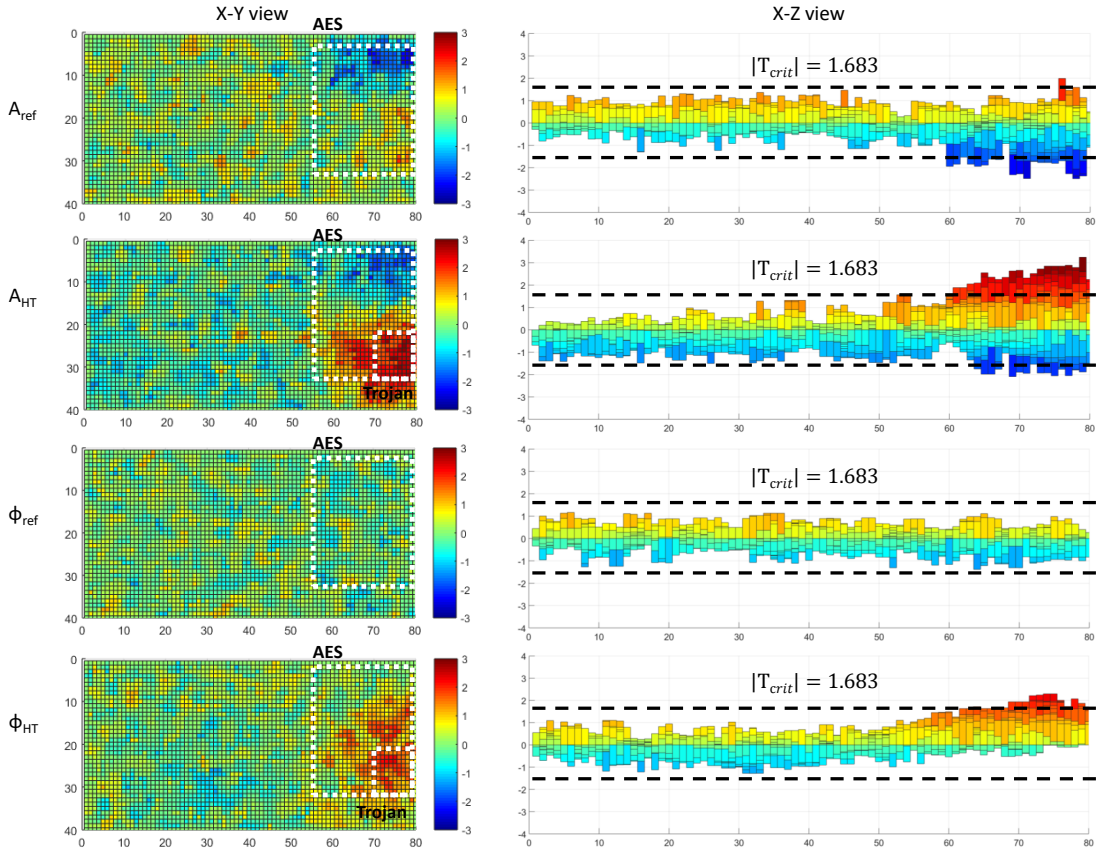

Fig. 2. Comparison between amplitude and phase results on Welch's T-test. A_{ref} and Φ_{ref} correspond to T-values obtained by application of Welch's T-test between to golden circuits for amplitude and phase measurements. A_{HT} and Φ_{HT} represent T-values obtained by application of Welch's T-test between a golden circuit and a HT infected circuit for amplitude and phase measurements. The dotted lines show the critical T-values (bilateral test) $\pm T_{crit}$

only 0.02 % of the Virtex 5 total area. Left images are shown in the x-y view in order to locate spatially the HT on the chip. The x-z view is presented on the right images.

The idea is now to determine if the T-values are high enough to indicate the presence of a HT. T-values represent how much the mean of the 100 acquired signals at a location (x,y) of the studied thermal map is different from the golden one. Hence, we can determine a threshold value, T_{crit}, so that any T-value (in absolute) exceeding this threshold can be considered as a sufficiently great difference in the thermal emission to be considered as a unusual thermal activity. For that we have found that the x-z view is of great help. Usually, T_{crit} is set so that 5 % error is tolerated when comparing samples of distributions with the same expectations. In our case, for 100 acquisitions, $T_{crit} = 1.683$.

Looking at A_{ref} in Fig. 2, one can observe that several dozens of coordinates fail the test ($|T| > |T_{crit}|$), even after standardization of the data. This indicates that something disturbs the measurements, rendering the means of the distributions at some coordinates different between the two golden ICs. Because the ICs are the same, one can think that the disturbance is coming from the room temperature variations which affect at the fourth power the measurements [6].

Nevertheless, one could observe that the resulting T-values are bounded between [-2.5; 2] when comparing the two golden ICs. These bounds can be used as thresholds to compare golden with suspect ICs. Indeed, when comparing a golden IC with an infected one, T-values rise well above 2, highlighting the presence of the stealthy HT. However such an approach remains valid if and only if environmental conditions (including the variance V(T) of temperature) are kept constant when comparing golden ICs with golden ones and golden ICs with suspect ones.

Considering now Φ_{ref}, one can observe that the obtained T-values exhibit the expected behavior. Indeed, they are centered symmetrically around 0 and enclosed in the [-1.683; 1.683] range, 1.683 being the value defining the critical area of the T-test for $\alpha = 0.05$ in our case. This is confirmed by Fig. 3 that shows:

- the cumulative distributions functions (CDF) of the student distribution with 198 degrees of freedom (100 samples in each sets of data)
- the empirical CDF (marked by '+') of the Φ_{ref} values
- the empirical CDF (marked by 'o') of Φ_{HT}

As presented in Fig. 3, Φ_{ref}'s CDF matches the Student distribution. This was confirmed by a goodness of fit test (two samples Kolmogorov-Smirnov test at $\alpha = 0.02$). On the contrary, Φ_{HT}'s CDF fails to pass the goodness of fit test and therefore does not match the Student distribution. This indicates the presence of the HT in the expected area of the IC. These results confirm that phase values are exempted of significant environmental or device linked variations affecting their distribution. Such results demonstrate the robustness of the proposed HT detection technique which is based on phase rather than amplitude values. They also demonstrate the benefits of considering phase rather than amplitude when dealing with low power ICs which can be explained by the great accuracy and stability of clocked IC operation in time.

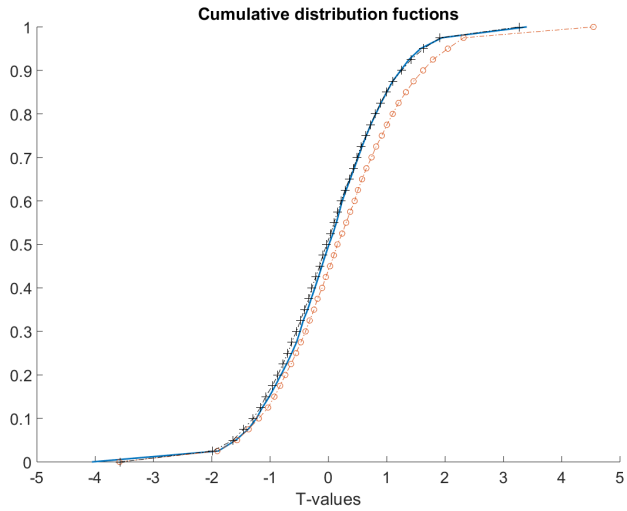

Fig. 3. Difference in cumulative distribution functions between the theoretical Student distribution (solid line), Φ_{ref} phase values ('+' marker) and Φ_{HT} phase values ('o' marker)

V. Conclusion

In this paper, we have presented an alternative way to investigate low thermal sources using IR signal phase measurements. We applied the latter improvements to Trojan identification and proved the efficiency of phase analysis. Obtained results demonstrate the superiority of single pixel detectors both in terms of detectivity and bandwidth. In the future, we believe that a higher sampling rate can greatly improve the spatial resolution and the detection limits of thermal maps.

References

[1] M. Lecomte, J. Fournier, and P. Maurine, An On-Chip Technique to Detect Hardware Trojans and Assist Counterfeit Identification, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 25, no. 12, pp. 3317-3330, Dec. 2017.

[2] L. Pyrgas, F. Pirpilidis, A. Panayiotarou, and P. Kitsos, Thermal Sensor Based Hardware Trojan Detection in FPGAs, 2017, pp. 268-273.

[3] J. Balasch, B. Gierlichs, and I. Verbauwhede, Electromagnetic circuit fingerprints for Hardware Trojan detection, 2015, pp. 246-251.

[4] X.-T. Ngo et al., Hardware Trojan Detection by Delay and Electromagnetic Measurements, 2015, pp. 782-787.

[5] M. Tehranipoor and F. Koushanfar, A Survey of Hardware Trojan Taxonomy and Detection, IEEE Design and Test of Computers, vol. 27, no. 1, pp. 10-25, Jan. 2010.

[6] Frank P. Incropera, David P. Dewitt, "Fundamentals of Heat and Mass Transfer", 5^{th} edition, pp. 700-746, 2001.

[7] R. S. Chakraborty, S. Narasimhan, and S. Bhunia. "Hardware Trojan: Threats and emerging solutions". In: High Level Design Validation and Test Workshop, 2009. HLDVT 2009. IEEE International, pp. 166-171, 2009.

[8] M. Cozzi, J.M. Galliere, P. Maurine, Thermal Scans for Detecting Hardware Trojan, Constructive side-channel analysis and secure design. New York, NY: Springer Berlin Heidelberg, 2018.