Machine Learning based Hardware Trojan Detection using Electromagnetic Emanation

NTT Secure Platform Laboratories

Junko Takahashi[1], Keiichi Okabe[1], Hiroki Itoh[1], Xuan-Thuy Ngo[2], **Sylvain Guilley**[2], Ritu-Ranjan Shrivastwa[2], Mushir Ahmed[2], Patrick Lejoly[2]

[1]NTT Secure Platform Laboratories, Tokyo, JAPAN
[2]Secure-IC, Cesson-Sévigné, FRANCE

ICICS 2020

August 26, 2020

Presentation Outline

1. Introduction
2. Contributions
3. Experiment Setup
4. Detection Results
5. Conclusion

Presentation Outline

1. Introduction
2. Contributions
3. Experiment Setup
4. Detection Results
5. Conclusion

Introduction

Outsourcing trend in Semiconductor

- Traditionally, Semiconductor industries designed and produced integrated chips by themselves
- Manufacturing techniques and standards evolved => they started to outsource the manufacturing step
- Semiconductor companies have also started to outsource the design and verification of their chips

New threats

- Counterfeit circuits
- Overproduction
- Reverse engineering => IP protection
- Hardware Trojan

Hardware Trojan

Definition

- Hardware Trojans are malicious modifications realized during the conception phase
- One inserted, they can realize critical damages such as:
  - Denial of Service
  - Degrade the performance of design
  - Change the behavior of the design
  - Leakage sensitive information
  - etc

HT Insertion

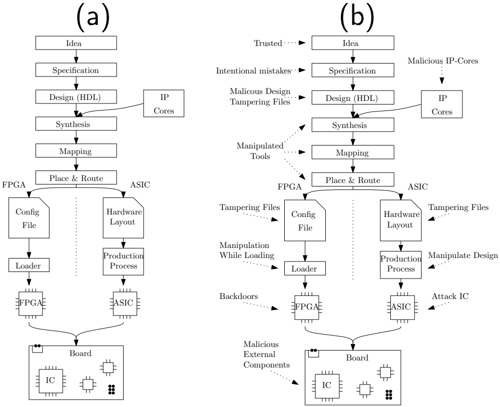
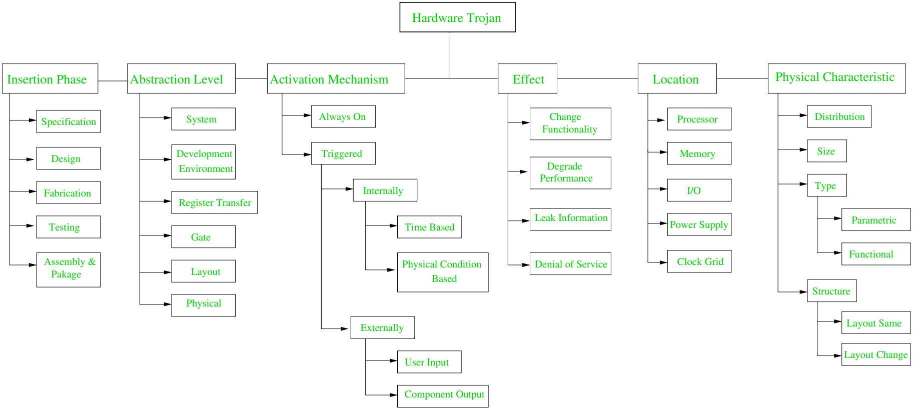


Figure 1: (a): IC development process. (b): HT scenario attacks on IC development process

HT Taxonomy



- Hardware Trojan
  - Insertion Phase
    - Specification
    - Design
    - Fabrication
    - Testing
    - Assembly & Pakage
  - Abstraction Level
    - System
    - Development Environment
    - Register Transfer
    - Gate
    - Layout
    - Physical
  - Activation Mechanism
    - Always On
    - Triggered
      - Internally
        - Time Based
        - Physical Condition Based
      - Externally
        - User Input
        - Component Output
  - Effect
    - Change Functionality
    - Degrade Performance
    - Leak Information
    - Denial of Service
  - Location
    - Processor
    - Memory
    - I/O
    - Power Supply
    - Clock Grid
  - Physical Characteristic
    - Distribution
    - Size
    - Type
      - Parametric
      - Functional
    - Structure
      - Layout Same
      - Layout Change

HT Examples

Year	Reporter	HTs detail
2018	Bloomberg	China used a tiny Chip to infiltrate 30 big U.S. Companies
2014	Defensenews.com	Specific US-made components designed to intercept the satellites' communications in France-UAE satellite
	Edward Snowden	NSA planted back-doors in Cisco products as routers
	Arstechnica and Spiegel	NSA secret toolbox used for inserting the backdoors and spy gadgets in different products
2012	Sergei Skorobogatov & Christopher Woods	The discovery of a backdoor inserted into the Actel/Microsemi ProASIC3 chips (military grade chip)
	Jonathan Brossard	A concept of a hardware backdoor called "Rakshasa" that China could embed in every computer
	Kryptowire	Found a backdoor on ZTE Android phones
From 2007	Academic	Many examples of HT on different targets (cryptography IPs, processors, Wireless etc.)

How to deal with this threat?

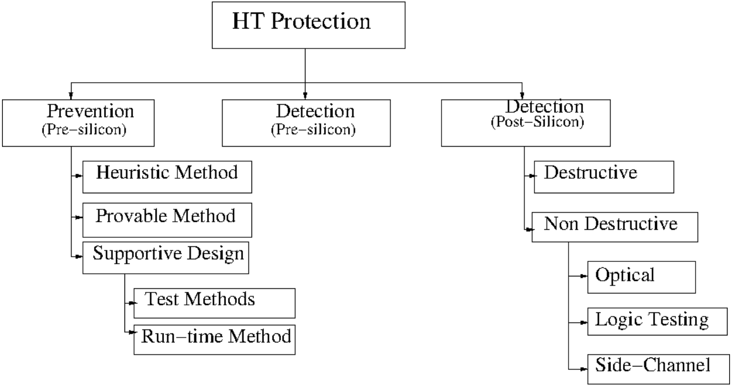


Figure 2: Possible countermeasures against hardware trojan horses

Presentation Outline

1 Introduction

2 Contributions

3 Experiment Setup

4 Detection Results

5 Conclusion

Contributions

In the state of the art

- Side-Channel methods seem to be good approaches for HT detection
- Electromagnetic based side channel method yields better precision
- Only some visual and statistical evaluation metrics are proposed:
    - Raw trace comparison
    - Using T-test coefficient for the detection

Some existing works:

- [BGV15], authors propose a detection method based on the comparison of T-test values. The experiment was done on an AES 128 bit design. However, in some cases, they can not detect the HT with a overhead smaller than 1.3%.
- [SKMH14], authors present the results of HT detection using the EM cartography traces. By comparing the averaged traces between the genuine and infected design, they can detect the HT only when the infected design is re-routed.
- [LJNM17], authors propose a detection methods by applying the One-Class SVM method on Side-Channel traces.
- Some other works based on EM traces are proposed. But, the detection performance is very low.

Contributions

In this paper, we

- Study the HT insertion on a generic processor
- Evaluate the performance of electromagnetic side-channel methods
- Propose new detection methods with a detection rate greater than 95% using:
  - Supervised machine learning (classification using both linear and non-linear approach)
  - Semi-supervised machine learning (an amalgamation of t-test and outlier detection algorithms)

Presentation Outline

1. Introduction
2. Contributions
3. Experiment Setup
4. Detection Results
5. Conclusion

Target designs

- For the experiment we need to prepare the test design with and without HT
- Target designs for the study:
  - PicoRV32 - Cliffordwolf
  - Freedom E300 - Sifive
- target boards
  - DE1 SoC with a Cyclone V Quartus FPGA for PicoRV32
  - Arty board with Arty Xilinx FPGA for Freedom E300

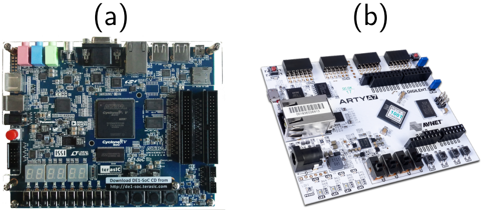


Figure 3: (a): DE1 SoC board (b): Arty board

HT insertion

HTs insertion for the experiment

- 2 HTs (HT1 & HT2) are inserted in PicoRV32 target
- HT1 & HT2 structure:
  - Trigger: Based on a specific DIV instruction (HT1) and registers (HT2)
  - Payload: Modify arbitrary the counter program
- 1 HT (HT3) is inserted in Freedom E300 target
- HT3 structure:
  - Trigger: Based on a specific DIV instruction
  - Payload: Modify arbitrary the privileged level of the processor

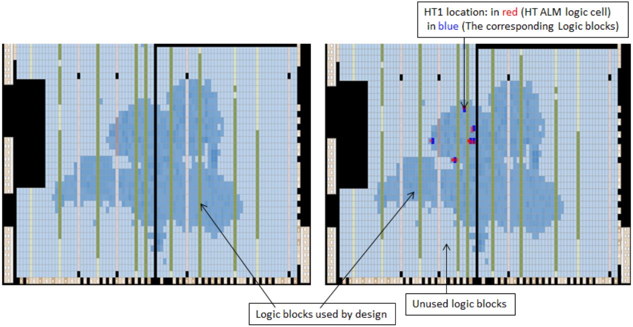
	Target design	Insertion phase	Overhead
HT1	PicoRV32	P&R	0.53%
HT2	PicoRV32	P&R	0.27%
HT3	Freedom	RTL	0.1%

Table 1: HT designs for the experimentation on RISC-V processors

HT insertion

HT insertion at P&R level

- The floorplan of the genuine and infected design is the same except where the HT is inserted
- Minimize the impact of the HT inserted in the target design
- This insertion is the worst case for the HT detection


HT1 location: in red (HT ALM logic cell)
in blue (The corresponding Logic blocks)
Logic blocks used by design
Unused logic blocks

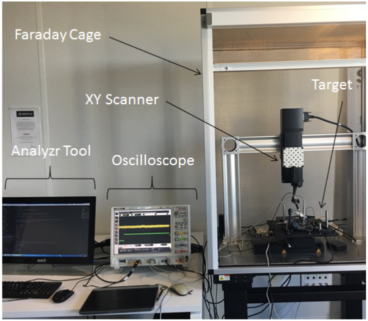
PicoRV32 without HT

PicoRV32 with HT1

Cartography platform

Cartography platform

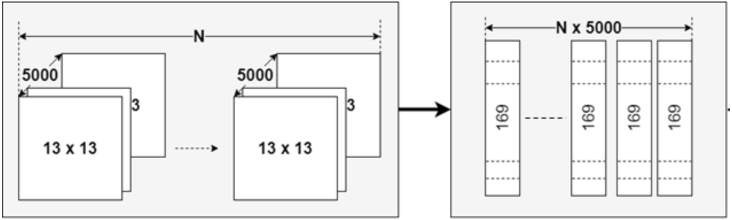
- An XYZ table is used to change the measurement position
- EM probe is used to measure the electromagnetic emanation of the target position
- An oscilloscope is used to measure and capture the EM traces
- Desktop computer is used to control and automate the process



Cartography platform

Cartography parameter

- One cartography consists of performing multiple measurements at several points on the target circuit.
- N_x steps of 2mm (for DE1 SoC board) and 1mm (for Arty-7 board) along X-axis and N_y steps (2mm for DE1 SoC) and 1mm for (Arty board) along Y-axis.
- For each measurement point, we have acquired N EM traces where N is the number of cartographies.
- Each EM trace contains T temporal samples ($T = 5000$ in this figure).



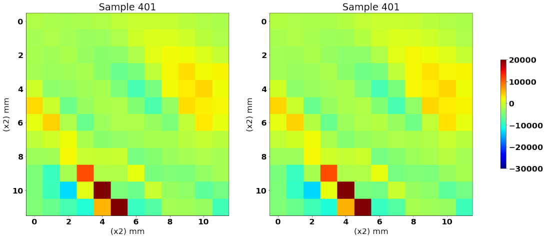
Presentation Outline

1. Introduction
2. Contributions
3. Experiment Setup
4. Detection Results
5. Conclusion

Detection method in the state of the art

Visual comparison

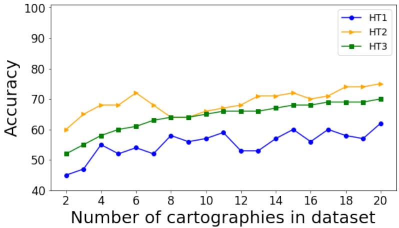
- Compare directly the raw traces between the design with and without HT
- Detect visually the difference created by the HT
- Difficult to detect in our case (where the HT is inserted at P&R level)

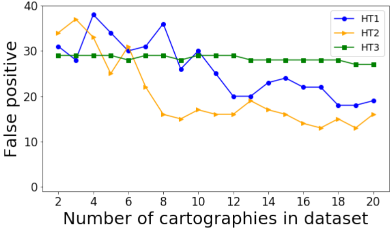


Detection method in the state of the art

T-test metric

- T-test is a metric used to detect if the mean of a population has a value specified in a null hypothesis or if the means of two different populations are equal.
- Calculate the T-test value between the test design and reference design
- Detection based on a threshold value
- This method has a poor performance
- It also depends of different parameters such as selected points and samples





Detection method in the state of the art

We notice that

- The existing methods can not detect our HTs.
- We need new and more sophisticated methods

We propose

- Two new detection methodologies:
  - Methodology 1: use the supervised machine learning algorithms
  - Methodology 2: use the combination of semi-supervised machine learning algorithms and T-test metric.

Proposed Detection Methodology

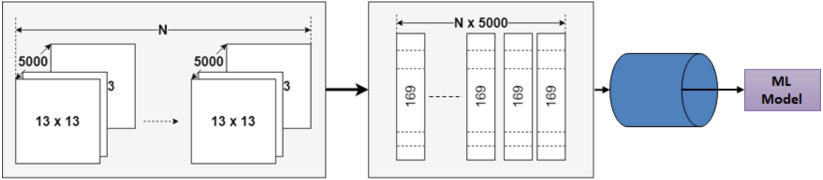
Methodology 1: Supervised Machine Learning methods

- Create a detection method with a dynamic threshold based on the training dataset
- Increase the performance
- The methodology is the following:
  1. Acquire the EM traces of reference design (EM_{ref}) and HT design (EM_{HT})
  2. Use these EM traces (EM_{ref} and EM_{HT}) to train the supervised machine learning algorithms
  3. Acquire the EM traces of test design EM_{test}
  4. Apply the trained models on EM_{test}, the models will decide if the test design is the same than reference or HT design
- Different supervised machine learning algorithms are applied:
  - Support Vector Machine (SVM)
  - Multi-Layer Perceptron (MLP)
  - Decision Tree Classification (DTC)
  - K-Nearest Neighbor Classifier (KNN)

Proposed Detection Methodology

Input data for Methodology 1

- Use raw EM trace for the training phase and detection phase
- 80% of EM traces of genuine and infected design are used for the training phase
- 20% of EM traces of genuine and infected design are used for the test phase
- Each input vector contains $N_x * N_y$ values which correspond to the raw EM all measurement point of one cartography and one sample



Detection results of Methodology 1

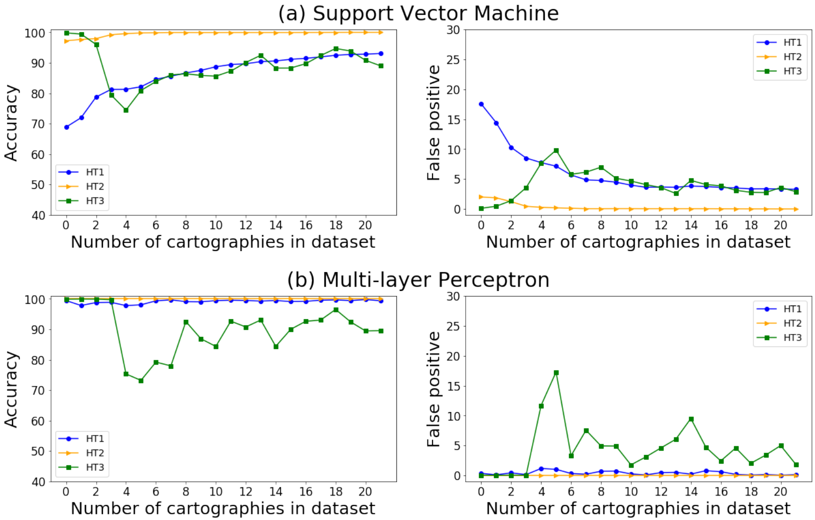
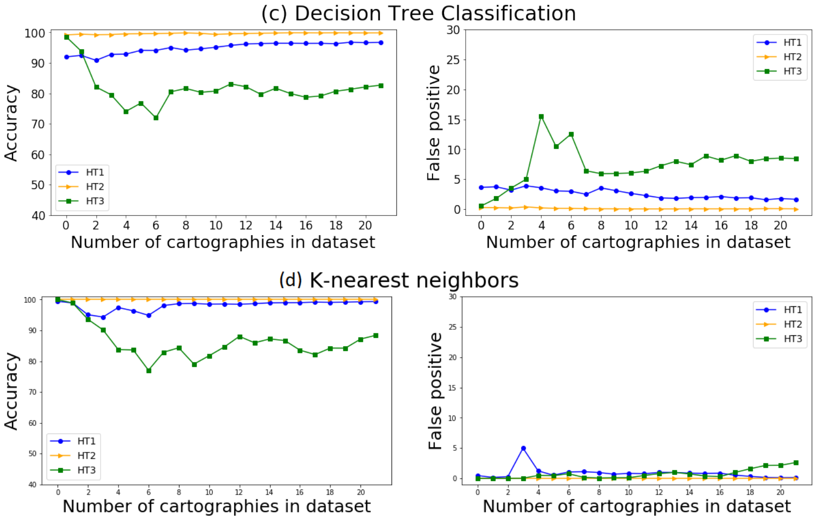


Figure 4: HTs detection using supervised machine learning algorithms (Part 1)

Detection results of Methodology 1



(c) Decision Tree Classification

(d) K-nearest neighbors

Figure 5: HTs detection using supervised machine learning algorithms (Part 2)

Proposed Detection Methodology

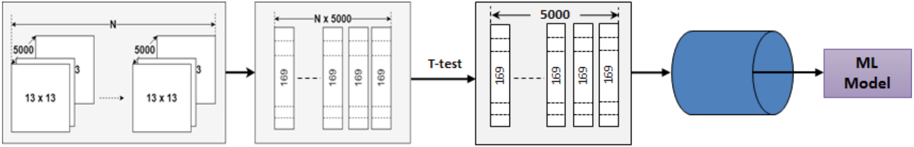
Methodology 2: Outlier/Novelty detection methods

- Create an automatic detection method without the HT dataset
- The methodology is the following:
  1. Acquire the EM traces of the reference design
  2. Compute the T-test value of the reference design (T_{ref})
  3. Train the Outliers detection algorithms using the T-test value (T_{ref})
  4. Acquire the EM trace of test design
  5. Compute the T-test value of the test design (T_{test})
  6. Test the trained Outlier detection algorithms with (T_{test}) to decide if the test design is the same (or not) than the reference design
- Different algorithms are applied:
  - One-Class Support Vector Machine (OCSVM)
  - Elliptical Envelope (EE)
  - Isolation Forest (IF)
  - Local Outlier Factor (LOF)

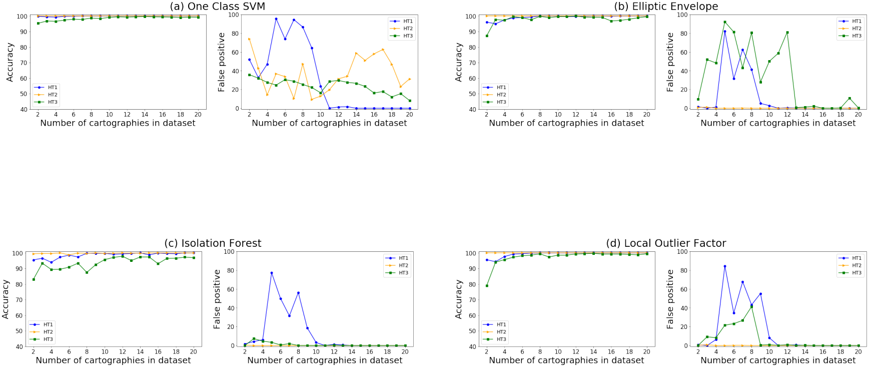
Proposed Detection Methodology

Input data for Methodology 2

- Use T-test coefficient of 50% of the genuine design dataset for the training
- Use T-test coefficient of 50% of the genuine design dataset and 100% of HT design dataset for the test phase
- Each input vector contains $N_x * N_y$ values which correspond to the T-test coefficient of all measurement point for one selected sample



Detection results of Methodology 2



(a) One Class SVM

(b) Elliptic Envelope

(c) Isolation Forest

(d) Local Outlier Factor

Figure 6: HTs detection using outliers detection

Result comparison with similar works

Inference on Methodology 2

- For the second method using the Outlier Detection algorithms we obtain promising results comparing to those in the state of the art as shown in table below.

	Method	Target	HT Size (%)	Detection rate
State-of-the-art	Raw trace comparison [SKMH14]	RISC-V	0.53, 0.27, 0.1	nc
	T-test [BGV15]	RISC-V	0.53, 0.27, 0.1	70%
	One-Class SVM [LJNM17]	RISC-V	0.53, 0.27, 0.1	60%
This paper	Supervised ML methods	RISC-V	0.53, 0.27, 0.1	≈ 90%
	Test & Outlier detection methods	RISC-V	0.53, 0.27, 0.1	≈ 100%

Presentation Outline

1 Introduction

2 Contributions

3 Experiment Setup

4 Detection Results

5 Conclusion

Conclusion

- HT is a real problem in the semiconductor domain.
- It can create a big problem for the industrial or governmental application.
- Different detection methods are proposed.
- Side-channel based methods seem to be a promising approach.
- We propose new methodology for the HT detection by combining EM side channel method, T-test processing and machine learning.
    - It has a greater performance (nearby 100% of detection rate) comparing to those in the state of the art
    - It propose a dynamic method
    - It does not need the dataset of HT for the training phase
- For the future work, we can have different interesting axes:
    - Create the simulated traces of the genuine design for the training phase
    - Evaluate the new methodologies for larger benchmarks (different target designs and different HTs)
    - Evaluate the performance of these new methodologies against the process variations

Thank you

References I

- J. Balasch, B. Gierlichs, and I. Verbauwhede, *Electromagnetic circuit fingerprints for hardware trojan detection*, 2015 IEEE International Symposium on Electromagnetic Compatibility (EMC), 2015, pp. 246–251.

- Y. Liu, Y. Jin, A. Nosratinia, and Y. Makris, *Silicon demonstration of hardware trojan design and detection in wireless cryptographic ics*, IEEE Transactions on Very Large Scale Integration (VLSI) Systems **25** (2017), no. 4, 1506–1519.

- O. Söll, T. Korak, M. Muehlberghuber, and M. Hutter, *Em-based detection of hardware trojans on fpgas*, 2014 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST), 2014, pp. 84–87.