A Post-Deployment IC Trust Evaluation Architecture

Yier Jin*, Dzmitry Maliuk† and Yiorgos Makris‡
*Department of Electrical Engineering and Computer Science, University of Central Florida
†Department of Electrical Engineering, Yale University
‡Department of Electrical Engineering, The University of Texas at Dallas
{yier.jin@eecs.ucf.edu, dzmitry.maliuk@yale.edu, yiorgos.makris@utdallas.edu}

***Abstract*—The use of side-channel parametric measurements along with statistical analysis methods for detecting hardware Trojans in fabricated integrated circuits has been studied extensively in recent years, initially for digital designs but recently also for their analog/RF counterparts. Such post-fabrication trust evaluation methods, however, are unable to detect dormant hardware Trojans which are activated after a circuit is deployed in its field of operation. For the latter, an on-chip trust evaluation method is required. To this end, we present a general architecture for post-deployment trust evaluation based on on-chip classifiers. Specifically, we discuss the design of an on-chip analog neural network which can be trained to distinguish trusted from untrusted circuit functionality based on simple measurements obtained via on-chip measurement acquisition sensors. The proposed method is demonstrated using a Trojan-free and two Trojan-infested variants of a wireless cryptographic IC design, as well as a fabricated programmable neural network experimentation chip. As corroborated by the obtained experimental results, two current measurements suffice for the on-chip classifier to effectively assess trustworthiness and, thereby, detect hardware Trojans that are activated after chip deployment.**

I. Introduction

Numerous hardware Trojan detection methods have been proposed to date, largely falling in two categories: *enhanced functional testing* and *side-channel fingerprint generation and checking*. The former are based on the assumption that infrequently occurring events will be employed by attackers to trigger the hardware Trojan, and therefore aim to include such events in the test plan [1]. The latter assume that a hardware Trojan will not alter the functionality but rather only the parametric profile of a chip. Therefore, they rely on a fingerprint constructed from parameters such as global power consumption [2], path delays [3], or currents on power grids [4], along with a trusted fingerprint region which is statistically learned from genuine circuits (golden models), in order to differentiate Trojan-infested from Trojan-free chips. Since the aforementioned methods are typically applied prior to chip deployment, a possible attack strategy to evade them is to design hardware Trojans that are dormant at test time and are only activated later in the field of operation. This can be easily achieved through a lapsed time or pre-specified input trigger [5]. Therefore, continuing to evaluate trustworthiness after chip deployment becomes equally important. To this end, we propose a general post-deployment trust evaluation architecture, which is based on on-chip measurement acquisition and classification, and we demonstrate its effectiveness on a wireless cryptographic IC.

II. Proposed Trust Evaluation Architecture

The proposed architecture for post-deployment trust evaluation is shown in Figure 1. The overall idea is fairly straightforward: after the circuit is deployed, the end-user can trigger the trust evaluation procedure at any time; during trust evaluation, on-chip resources are used to apply a known stimulus to the circuit and to obtain parametric measurements, which are then assessed on-chip to decide whether the circuit is operating within a trusted region. To this end, several components are added to the chip, along with the original circuit:

- A programmable on-chip non-volatile stimulus storage component (i.e., Flash, EEPROM, or OTPROM) and a multiplexer through which the known necessary excitation stimulus is provided to the circuit.
- Measurement acquisition sensors, to obtain the parametric signature of the circuit in response to the known stimulus.
- An on-chip classifier, to assess the parametric signature obtained via the sensors and to decide whether the circuit operation is trusted or not.
- Programmable on-chip non-volatile storage for programming the topology and the weights that define the region accepted as trusted by the classifier.

We point out that programmability and non-volatility are required, so that the actual stimulus, the topology of the classifier, and the region accepted as trusted are stored on the chip only *after* it is fabricated. Thereby, a potential attacker is not privy to this information. While the attacker may be able to understand what parameters are being measured, without knowledge of the stimulus, the actual structure of the classifier and the definition of the trusted region, it will be very difficult to design a hardware Trojan that evades detection. In essence, the proposed architecture counteracts the element of surprise possessed by the attacker (i.e., the ability to choose the location, functionality, and time of activation of the hardware Trojan) by a similar element of surprise possessed by the defender (i.e., the ability to choose the type of parametric signature, the method and bounds for assessing its trustworthiness, and the time of trust evaluation).

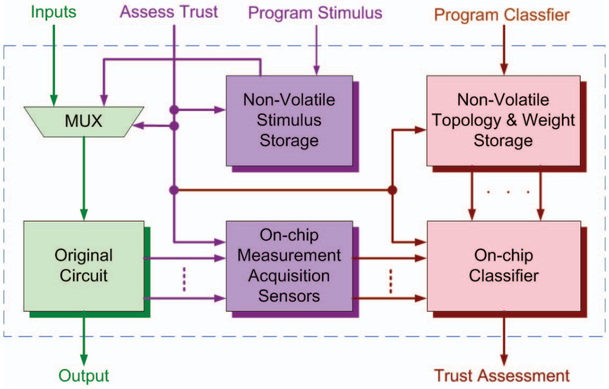

Fig. 1. Proposed post-deployment trust evaluation architecture

III. Experimental Platform

A. Target Circuit

The experimental platform used to demonstrate effectiveness of the proposed post-deployment Trojan detection method is based on the mixed-signal wireless cryptographic IC described in [6]. This chip takes plain-text at its input, encrypts it using an on-chip stored key, and then transmits the cipher-text on a public wireless channel. The basic platform architecture is divided into three parts: (i) the digital part, which includes a pipelined Digital Encryption Standard (DES) core, an output buffer, and a serializer serving as the interface between the digital and analog parts, (ii) the analog part, which is an ultrawide-band (UWB) transmitter, and (iii) the on-chip trust evaluation resources. These include an on-chip non-volatile serial-in parallel-out 64-bit register to hold the trust evaluation stimulus, two Built-in Current Sensors (BICS)along with envelop detectors and DC-DC converters to obtain the side-channel fingerprint of the chip, and a neural network to classify it as trusted or untrusted. Our current experimentation platform consists of SPICE-level simulation models for all components, except for the neural classifier. The latter is emulated through a programmable analog neural network experimentation chip.

B. On-Chip Trust Evaluation Resources

The on-chip trust evaluation part performs two tasks, namely parametric measurement acquisition and data classification. Parametric measurements are obtained via on-chip sensors in response to a known stimulus, which is also stored on-chip using a non-volatile serial-in parallel-out (SIPO) shift register. A `BIST_in` signal is used to fill in the 64-bit wide register with a value *after* fabrication and prior to deployment. Another `BIST_en` signal controls the data flow to the digital/analog interface. When `BIST_en` is ‘0’, the input of the interface is the ciphertext to be sent by the UWB transmitter while when it is ‘1’, the pattern stored in the SIPO register is sent to the UWB transmitter, in order to perform trust evaluation. In this work, we use two current measurements obtained from the UWB transmitter for trust evaluation. The output of the BICS is a high frequency signal which we convert to a DC voltage through a CMOS envelope detector. Both the current sensor and envelope detector are CMOS designs so that they are compatible with other parts of the circuit. A DC-DC converter is then used to match the measurement to the input range of the neural network which will perform data classification.

C. On-Chip Classifier

To demonstrate in silicon that an on-chip classifier can, indeed, detect a hardware Trojan upon its activation in the operation field, we employ an analog neural network experimentation chip. Using this programmable chip, we implement artificial neural networks, which we then train to learn (through a training set of chips) the mapping between the current measurements obtained from the two BICS which we integrated inside the UWB transmitter, and the trusted operation region. The trained neural networks can then be evaluated with respect to their capability to detect Trojan-infested chips using a validation set. We note that an analog VLSI implementation of the neural classifier is necessary in order to contain the area and power overhead of the proposed trust evaluation.

TABLE I
Type I Trojan Classification

		Classified by hardware		Classified by software	
		Dormant	Activated	Dormant	Activated
Actual	Dormant	99.9%	0.1%	100%	0%
	Activated	2.8%	97.2%	1.3%	98.7%

TABLE II
Type II Trojan Classification

		Classified by hardware		Classified by software	
		Dormant	Activated	Dormant	Activated
Actual	Dormant	99.8%	0.2%	100%	0%
	Activated	0%	100%	0%	100%

D. Hardware Trojans

In addition to the Trojan-free circuit, two alternative hardware Trojan-infested variants of the wireless cryptographic IC are also designed. These are of similar structure and working principle to the Trojans we introduced in [6] with the exception that both Trojans are dormant during the testing stage and are only activated after deployment. Through simple modifications on only the digital portion of the chip, they leak the encryption key by hiding it in the wireless transmission amplitude (Type-I) or frequency (Type-II) margins allowed due to process variations; thus, they ensure that the circuit continues to comply to its functional specifications and, thereby, evade testing both on the digital and on the analog side.

IV. Experimental Results

We first train the on-chip classifier with data from Monte-Carlo generated Trojan-free chip instances. Then, we assess its effectiveness in correctly classifying the two types of Trojan-infested chip populations. In order to obtain a global picture, we present the trained classifier with the data from both when the Trojan is dormant and when the Trojan is activated. Tables I and II report the confusion matrices for the Type-I and Type-II Trojan-infested chip populations, respectively. For comparison, the effectiveness of the software version of the classifier is also reported, demonstrating that the error due to the hardware implementation is minimal. While not zero, the false positive and false negative rates are very low, indicating that the proposed method provides an effective post-deployment trust evaluation capability.

References

[1] F. Wolff, C. Papachristou, S. Bhunia, and R. S. Chakraborty, “Towards Trojan-free trusted ICs: Problem analysis and detection scheme,” in *IEEE Design Automation and Test in Europe*, 2008, pp. 1362–1365.

[2] D. Agrawal, S. Baktir, D. Karakoyunlu, P. Rohatgi, and B. Sunar, “Trojan detection using IC fingerprinting,” in *IEEE Symposium on Security and Privacy*, 2007, pp. 296–310.

[3] Y. Jin and Y. Makris, “Hardware Trojan detection using path delay fingerprint,” in *IEEE International Workshop on Hardware-Oriented Security and Trust*, 2008, pp. 51–57.

[4] R. Rad, J. Plusquellic, and M. Tehranipoor, “Sensitivity analysis to hardware Trojans using power supply transient signals,” in *IEEE International Workshop on Hardware-Oriented Security and Trust*, 2008, pp. 3–7.

[5] Y. Jin, N. Kupp, and Y. Makris, “Experiences in hardware Trojan design and implementation,” in *IEEE International Workshop on Hardware-Oriented Security and Trust*, 2009, pp. 50–57.

[6] Y. Jin and Y. Makris, “Hardware Trojans in wireless cryptographic ICs,” *IEEE Design and Test of Computers*, vol. 27, pp. 26–35, 2010.