Hardware Trojans in Wireless Cryptographic ICs: Silicon Demonstration & Detection Method Evaluation

Yu Liu*, Yier Jin†, and Yiorgos Makris*
*Department of Electrical Engineering, The University of Texas at Dallas
†Department of Electrical Engineering and Computer Science, University of Central Florida

***Abstract*—We present a silicon implementation of a hardware Trojan, which is capable of leaking the secret key of a wireless cryptographic integrated circuit (IC) consisting of an Advanced Encryption Standard (AES) core and an Ultra-Wide-Band (UWB) transmitter. With its impact carefully hidden in the transmission specification margins allowed for process variations, this hardware Trojan cannot be detected by production testing methods of either the digital or the analog part of the IC and does not violate the transmission protocol or any system-level specifications. Nevertheless, the informed adversary, who knows what to look for in the transmission power waveform, is capable of retrieving the 128-bit AES key, which is leaked with every 128-bit ciphertext block sent by the UWB transmitter. Using silicon measurements from 40 chips fabricated in TSMC's $0.35\mu m$ technology, we also assess the effectiveness of a side channel-based statistical analysis method in detecting this hardware Trojan.**

I. Introduction

Hardware Trojans are malicious modifications introduced in a manufactured IC, which can be exploited by a knowledgeable adversary to cause incorrect results, steal sensitive data, or even incapacitate a chip [1], [2]. The problem of hardware Trojans has recently caught the attention of multiple governments and industry across the globe, who are realizing the repercussions of inadvertent deployment of hardware Trojan-infested ICs in sensitive applications and are investing in understanding the risk and developing appropriate solutions. Indeed, traditional IC test methods fall short in detecting hardware Trojans, as they are mainly geared towards identifying modeled defects; therefore, they cannot reveal unmodeled malicious inclusions, especially when the latter are carefully hidden and do not visibly alter the functionality of the IC.

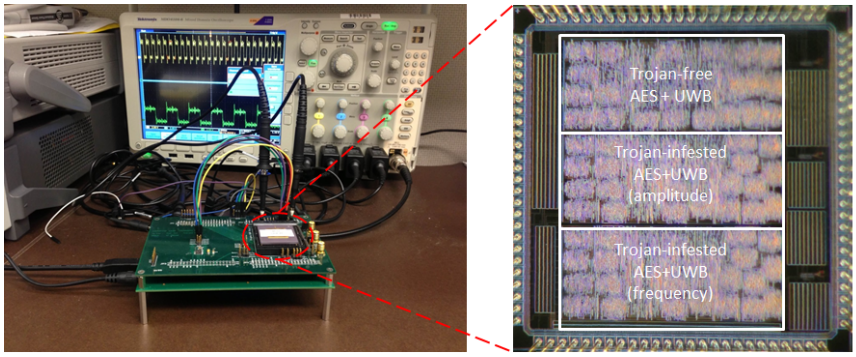
Among the various hardware Trojan detection methods proposed by researchers over the last few years, statistical analysis of side-channel measurements has received the lion's share of attention. The underlying premise of this approach is that hardware Trojans will distort the side-channel parametric profile of an IC, even if they do not alter its functionality. While for a well-designed hardware Trojan this distortion is minute and carefully hidden within the design margins allowed for process variation, it is systematic; therefore, statistical analysis should be able to identify the presence of additional structure in the side-channel parametric profile of an IC and, thereby, reveal its presence. Accordingly, assuming availability of a small, representative set of trusted Trojan-free ICs, classifiers can be trained to discern between Trojan-free and Trojan-infested chips.

Starting with the global power consumption-based method presented in [3] and the path delay-based method introduced in [4], constructing *fingerprints* of ICs based on side-channel parameters and using these fingerprints to statistically assess whether an IC is contaminated by a hardware Trojan or not became a popular direction. Indeed, numerous researchers in the hardware trust area developed this idea further by using various side-channel measurements, including power supply transient signals [5], [6], leakage currents [7], regional supply currents [8], and temperature [9], as well as multi-parameter combinations thereof [10], [11]. While all of these methods targeted digital circuits, a similar method using side-channel fingerprinting to detect hardware Trojans in analog/RF ICs, and more specifically, in wireless cryptographic ICs was also proposed in [12]. As pointed out therein, the analog/RF domain is an attractive attack target, since the wireless communication of these chips with the environment over public channels simplifies the process of staging an attack without obtaining physical access to the I/O of the chip. On the other hand, signals in an analog/RF IC are continuous and highly-correlated to one another; hence, the likelihood of a modification disturbing these correlations is very high. As a result, side channel-based hardware Trojan detection methods are very effective in this domain, as shown using a Trojan-free and two Trojan-infested versions of a wireless cryptographic IC in [12].

The work presented herein seeks to corroborate the findings of [12] through actual silicon measurements, as opposed to simulation-based results. Indeed, silicon measurements are essential in order to convincingly assess effectiveness of side-channel fingerprinting methods, especially in the analog/RF domain. To this end, we follow the same general principles for leaking secret information and the same hardware Trojan detection method, as introduced by the authors of [12], although our design is slightly different. To our knowledge, this is the first silicon demonstration of a working hardware Trojan in a wireless cryptographic IC and the first evaluation of side channel-based statistical analysis methods for detecting it.

The remainder of this paper is structured as follows. In Section II, we introduce the chip that we designed and fabricated for the purpose of this study. Specifically, we describe the Trojan-free and the Trojan-infested versions of an AES+UWB wireless cryptographic IC, the method through which the key is leaked by it, as well as issues pertaining to the robustness of the attack. Then, in Section III, we discuss how the hardware Trojan evades detection by traditional manufacturing test methods. Finally, in Section IV, we assess the effectiveness of a side-channel method in revealing the presence of a hardware Trojan based on statistical analysis of transmission power.

978-1-4799-1071-7/13/$31.00 ©2013 IEEE



Technology	TSMC 0.35 µm
Chip Size	3mm x 3mm
Modulation Scheme	FSK-OOK
Data Rate	Up to 96MHz
Pulse Width	7ns to 48ns
Central Frequency of '0'	900 MHz
Central Frequency of '1'	1.6 GHz

Fig. 1. Experimental platform, die photograph, and circuit specifications

II. Wireless Cryptographic IC

The wireless cryptographic circuit used in this study has two main parts, a digital and an analog. The digital part consists of an Advanced Encryption Standard (AES) core and an output buffer. The analog part is an Ultra-Wide-Band (UWB) transmitter, which is very small and easy to integrate on-chip. On our experimental platform, which is shown in Fig. 1, there are three versions of the circuit integrated on the same die, a Trojan-free and two Trojan-infested. The Trojan-infested circuits leak the AES encryption key by hiding it in the wireless transmission power amplitude/frequency[1] margins allowed for process variations, while ensuring that the circuit continues to meet all of its functional specifications.

The chip was designed and fabricated in TSMC's $0.35\mu m$ process, with all 40 chips received functioning correctly. The digital part can run at a frequency of up to 48MHz and the UWB transmitter has a data rate of up to 96Mbps. Specifications of the chip are also listed in Fig. 1. The packaged die sits in a socket of a custom PCB, which is connected to an Opal Kelly XEM 3010 FPGA board, through which the wireless cryptographic IC can be controlled from a PC via Matlab. The bias voltage of the UWB transmitter is controlled by an 8-bit DAC (AD5668) on the PCB.

A. Trojan-free Version

A system-level block diagram of the circuit is shown in Fig. 2. The AES core receives plaintext in blocks of 128 bits, which it encrypts using a 128-bit key that is stored on-chip through the 'key' input. The width of the encryption key determines the number of transformation rounds to which the plaintext is subjected during encryption. In this case, after 10 rounds of transformation, the plaintext is encrypted into ciphertext, which is stored in an output buffer in blocks of 128 bits, until it is transmitted. The UWB transmitter designed in this platform includes a baseband pulse generator and an RF pulse generator, as shown in Fig. 3. In our design, frequency-shift keying (FSK) is used to distinguish the polarity of a bit, while on-off keying (OOK) is used to separate adjacent bits. Bit values of '0' and '1' are separated and converted to RZ (return-to-zero) format in the baseband pulse generator. The pulse width is controlled

[1]Due to space limitations, we only report results from the first Trojan (amplitude-based) herein; results using the second hardware Trojan (frequency-based) are similar.

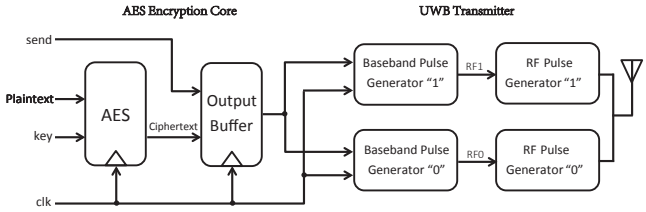


Fig. 2. System-level block diagram

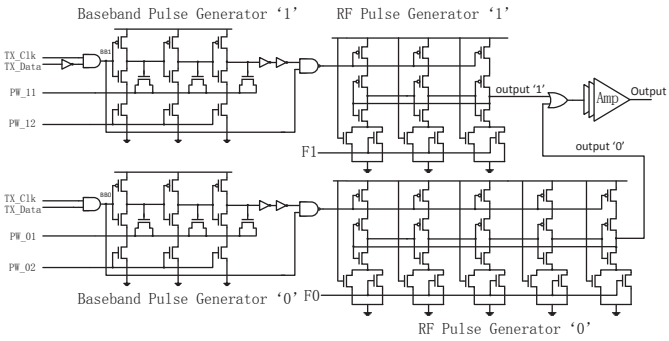


Fig. 3. UWB transmitter schematic

by two PW signals. The output of the baseband pulse generator controls the input of the RF Pulse Generator. Here, a ring oscillator is used to generate the RF pulse, and the pulses of signals '0' and '1' are assigned to two different frequencies. Signals F0 and F1 are used to control the pulse frequency. The modulation waveform of the UWB transmitter is shown in Fig. 4. An example of a typical transmission of a '1' and '0' is shown in Fig. 5. We note that transmission of signal '1' has higher frequency and lower amplitude than transmission of signal '0'.

B. Trojan-infested Version

Hardware Trojans leverage the fact that the underlying hardware modifications needed are very simple. In our experimental platform, minor additions to the digital and analog part of the circuit are required to leak the encryption key over the public channel, as shown in Fig. 6. Specifically, on the digital side, the designed hardware Trojan taps into the register that stores the 128-bit AES key and steals one bit at a time. The value of the stolen key bit is passed to the UWB transmitter, where it used to control the wireless transmission

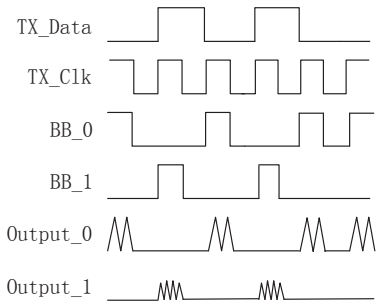


Fig. 4. Modulation waveform

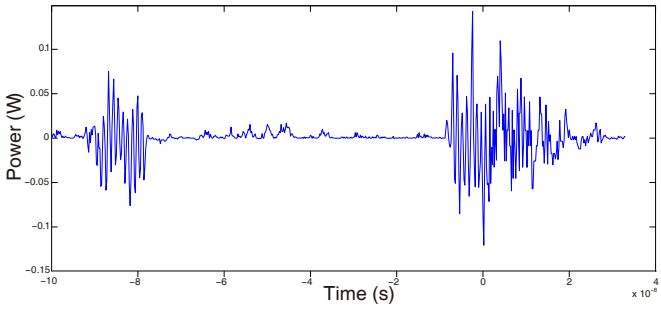


Fig. 5. Transmission power waveform while sending '1' and '0'

amplitude during the transmission of one ciphertext bit. Overall, along with every 128-bit block transmitted by the UWB transmitter, the 128-bit key is also leaked. On the analog side, the modification needed to leak a stolen key bit with each transmitted ciphertext bit is also very simple, leveraging the design margins provided to account for fabrication process variation. Specifically, a PMOS transistor is added to the output of the power amplifier (PA) of the UWB transmitter, and the stolen key bit is connected to the gate of the PMOS transistor. Accordingly, when the stolen key bit is '0', the PMOS transistor is turned on and draws a small additional current from the power supply to the output, thereby slightly increasing the transmission power. Conversely, when the stolen key bit is '1', the PMOS transistor is turned off, so no additional current is drawn to the output.

Fig. 7 shows the impact of the introduced hardware Trojan on the transmission power waveform of a Trojan-infested chip. Fig. 7(a) contrasts the power waveforms for transmitting a logic '0' when the stolen key bit is '1' and '0', respectively. In the latter case, the slight increase in transmission power is evident across the waveform, with the difference peaking at 14mW. Similarly, Fig. 7(b) contrasts the power waveforms for transmitting a logic '1' when the stolen key bit is '1' and '0', respectively, with the difference in transmission power peaking at 8mW. We emphasize that this slight increase in transmission power when the PMOS transistor is turned on (i.e. when the stolen key bit is '0') is very small and leaves the circuit well within its functional specification margins allowed for process variations and operating condition fluctuations. In other words, all of these transmissions appear perfectly legitimate and do not raise any suspicions, as they could have been the outcome of a chip from the Trojan-free distribution.

C. Stealing the Key

Despite being hidden in the process variation margins, the impact of the hardware Trojan on the transmission power waveform suffices for the informed adversary to obtain the secret key and, by extension, the plaintext by deciphering the ciphertext. We note that the attacker does not need to know the exact shape of the waveform when a key of value '0' and a key of value '1' is leaked. In fact, it is impossible to know this information, since every chip will be affected differently by process variations. Indeed, the attacker does not rely on absolute values. Rather, it is the minute relative difference between transmissions by the same chip that gives away the secret. All the attacker has to do is listen to the public wireless transmission channel, focusing on the parameter manipulated by the hardware Trojan (i.e. amplitude), in order to observe the different levels which correspond to a key bit of '1' and '0' respectively, when a ciphertext bit of value '0' and a ciphertext bit of valure '1' is transmitted (i.e. the waveforms of Fig. 7). Once these four waveforms are known to the attacker, observing the transmission of a 128-bit block suffices to obtain the entire 128-bit AES key.

Fig. 8 shows an example of how the encryption key is leaked. This example zooms in on an 8-bit portion of a 128-bit ciphertext transmitted by the UWB transmitter. The value of this 8-bit snippet is '00110011' and the corresponding 8-bit key portion is '10101010'. Each bit transmission in this example is perfectly legitimate, within the specifications of the circuit. However, comparative observation of the power transmission amplitude in the waveforms of the transmitted bits reveals the value of the key bits to the attacker.

D. Attack Robustness

For the designed hardware Trojan to facilitate a robust attack, the difference between the transmission power waveforms when the stolen key bit is '0' and '1' should be discernible even in the presence of measurement noise and environmental variations. When measuring transmission power, unavoidable measurement noise is introduced due to the accuracy of the test equipment (i.e. starting point and step size precision), resulting in slightly different outcomes for the same waveform. Environmental conditions such as temperature, EMI, and test-board setup may also impact the measurements. To assess the attack robustness of our hardware trojan to noise, we conducted 10 repetitions of the same measurements at different times in different locations and under slightly different temperature conditions. Fig. 9(a) shows the 10 power waveforms obtained when transmitting a '1' and a '0' when the stolen key bit is a '0', while Fig. 9(b) shows the same measurements when stolen key bit is '1'. As may be observed, the lowest peak amplitude among the 10 repetitions shown in Fig. 9(a) is always above 60mW and 80mW for transmitting a '1' and a '0', respectively, when the key bit is '0', while the corresponding highest peak amplitude shown in Fig.9(b) never exceeds these values when the key bit is '1'. Hence, the difference is clearly distinguishable and the hardware Trojan can robustly leak the encryption key.

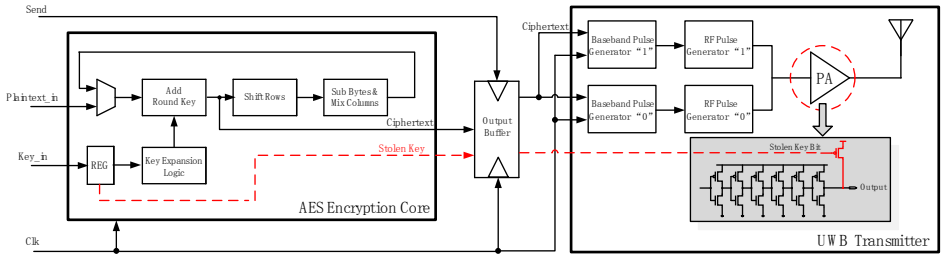


Fig. 6. Hardware Trojan modifications in digital and analog parts

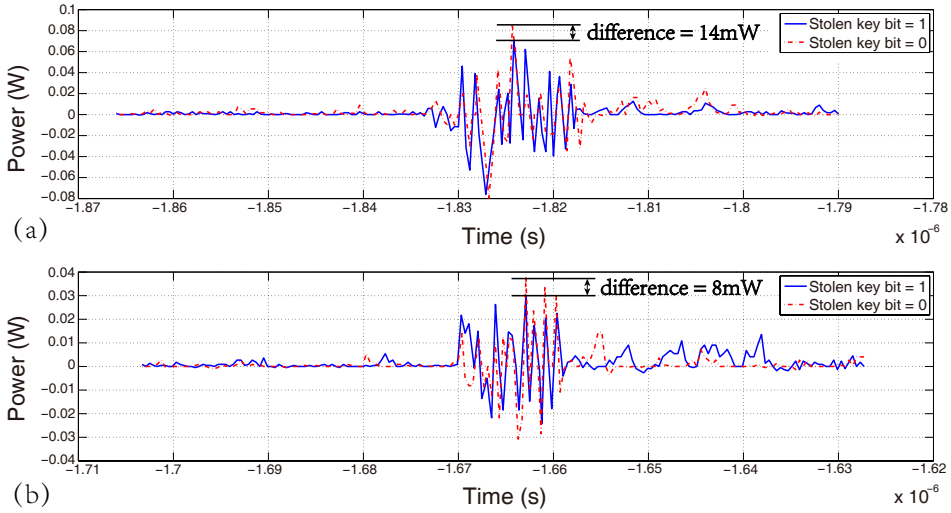


Fig. 7. Difference in transmission power waveform of Trojan-infested chip when the stolen key bit is '0' and '1' while transmitting (a) a ciphertext bit of value '0', and (b) a ciphertext bit of value '1'.

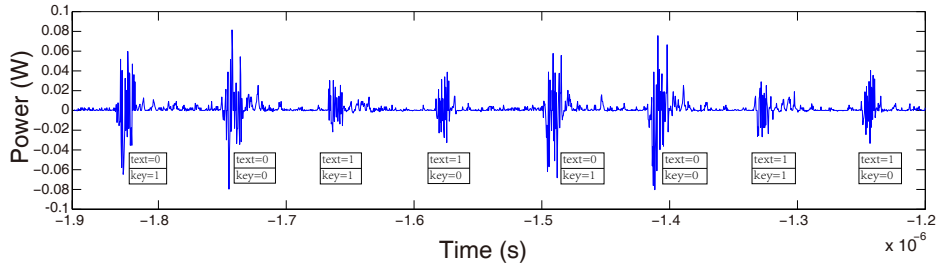


Fig. 8. Transmission power waveform of a 8-bit block

III. Detection Evasion

The hardware Trojan introduced in our design does not alter the functionality of the AES circuit, as it only taps into the register holding the encryption key. No functional or structural digital test targeting stuck-at or transition faults is, therefore, going to expose it. Also, the added capacitive load for leaking the key, one bit at a time, is very low to make the circuit fail any delay tests or to be picked-up by statistical hardware Trojan detection methods, such as [4]. Similarly, on the UWB side, the impact of the introduced hardware Trojan (i.e. PMOS transistor) is hidden within the process variation margins. In other words, for the vast majority of fabricated devices, the transmission power waveform will continue to be within the UWB specifications. It is possible, however, that for a very small number of chips at the tails of the distribution, the extra nudge provided by the Trojan might push them outside the specifications, thereby slightly reducing yield. Nevertheless, such yield loss could be caused by many other reasons (process drifts, material impurities, mask misalignment, measurement noise, etc.) and there is no way to attribute it to the presence of a hardware Trojan. In our case, none of the 40 Trojan-infested chips ended up outside the specifications, while all of them could robustly leak the secret key. System-level test is also not going to reveal this hardware Trojan, since it does not transmit any additional information and it does not violate the transmission protocol in any way.

To demonstrate the difficulty in detecting this hardware Trojan, in Figs. 10(a) and (b) we show the measured transmission power for transmitting a '0' and a '1' by each of the 40 hardware Trojan-free and hardware Trojan-infested circuits on the 40 fabricated chips. Each of the distributions is enclosed in the $\mu \pm 3\sigma$ transmission power envelop of the hardware Trojan-free circuits. The key observation based on this figure is that the two distributions are very similar. Clearly, given any one of these 80 transmission power waveforms, it is very difficult,

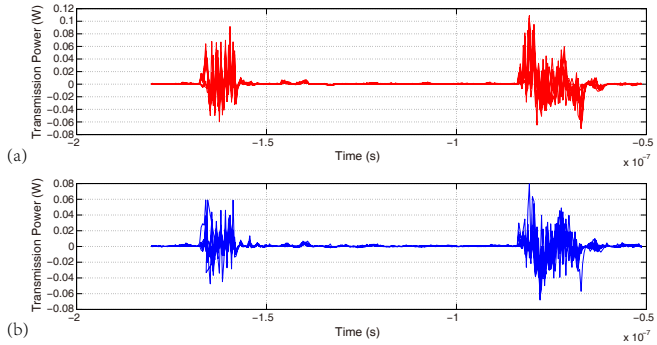


Fig. 9. 10 repetitions of measuring transmission power when transmitting a '0' and a '1' when the stolen key bit is (a) '0', and, (b) '1'.

if not impossible, to definitively tell whether it comes from a hardware Trojan-free or a hardware Trojan-infested circuit.

IV. Side-Channel Detection Method Evaluation

As mentioned in Section I, the underlying premise of statistical side channel-based detection methods is that the distortion imposed by hardware Trojans on the parametric profile of an IC is *systematic*, even though it is hidden within the design margins allowed for process variations. For example, the hardware Trojan introduced herein increases slightly the transmission amplitude when the stolen key bit value is '0', without violating any transmission specifications. This systematic impact of the attack is indispensable, since the adversary relies on it in order to discern the hidden information. However, any systematic component, subtle as it might be, imposes added statistical 'structure' to the transmission power of a population of chips. This added 'structure' is precisely what statistical side channel-based hardware Trojan detection methods rely on. Let us demonstrate this point using the silicon measurements from our chips.

We randomly selected six different blocks of plaintext, which we encrypted through the AES using a randomly chosen 128-bit key. Each of the resulting six blocks of ciphertext was then transmitted by the UWB transmitter and the total transmission power for each block over the public channel was measured for each of the 40 Trojan-free and 40 Trojan-infested circuits. In Fig. 11(a), we project these populations to a randomly chosen subset of three out of these six measurements. Evidently, the populations fall upon each other and cannot be separated in this space through simple upper/lower limits on each axis (i.e. through a hyper-cube). This remains the case for all other subsets of three out of the six measurements. This is expected, as the transmission power for each individual block remains within the acceptable specification margins for all of these circuits. In other words, by simply examining transmission power of blocks by individual chips, it is not possible to reveal the presence of a hardware Trojan.

However, when we perform even a very simple statistical processing of the same information (i.e. the total transmission power for transmitting the same six ciphertext blocks as above) from all the chips, such as Principal Component Analysis (PCA) [13], things start to become very interesting. In Fig. 11(b), we project again the two populations on the three principal components of this data. Evidently, in this space, the two populations are clearly separable. For example, in the figure we show how a simple Minimum Volume Enclosing Ellipsoid (MVEE) [14] can be used to enclose the Trojan-free chips and serve the purpose of a classifier, which decides whether a chip is Trojan-free (i.e. inside the MVEE) or Trojan-infested (i.e. outside the MVEE). For our experiment, it can be observed that this simple classification method would incur no false alarms, demonstrating the effectiveness of statistical side-channel hardware-Trojan detection in this case.

We emphasize that in training the classifier we only use measurements from the Trojan-free chips. No assumptions are made and no knowledge is used regarding the functionality or the design of the hardware Trojan. Therefore, this method should remain equally effective for any Trojan that manipulates transmission power to leak information.

V. Conclusion

Wireless cryptographic ICs provide a tangible objective and constitute an attractive target for hardware Trojans. Not only do these ICs hold valuable secret information, but also they communicate over public channels, thereby simplifying the attack. Indeed, leaking the secret key by hiding it in the wireless transmission power, to which an adversary has access, is fairly straightforward and requires very little circuit modification. More importantly, this can be done without violating any digital, analog, or system-level specifications, rendering traditional test methods ineffective in detecting such hardware Trojans, the impact of which is carefully concealed within the design margins allowed for process variations. In this sense, transmission by a hardware Trojan-infested wireless cryptographic IC appears perfectly legitimate and, in isolation, cannot be differentiated from that of a hardware Trojan-free chip. Nevertheless, due to the systematic nature of the hardware Trojan impact, statistical analysis is capable of revealing the presence of a hardware Trojan, without requiring any a priori knowledge about the particulars of the attack. The above observations were demonstrated using 40 chips from a wireless cryptographic IC design, consisting of an AES encryption core and a UWB transmitter, which we designed and fabricated in TSMC's $0.35\mu m$ process. To our knowledge, this is the first silicon demonstration of hardware Trojans in wireless cryptographic ICs and the first evaluation of the popular side channel-based detection method using actual measurements.

References

[1] S Adee, "The hunt for the kill switch," May 2008.
[2] R. Karri, J. Rajendran, K. Rosenfeld, and M. Tehranipoor, "Toward trusted hardware: Identifying and Classifying Hardware Trojans," *IEEE Computer Magazine*, Oct. 2010.
[3] 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.
[4] 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.

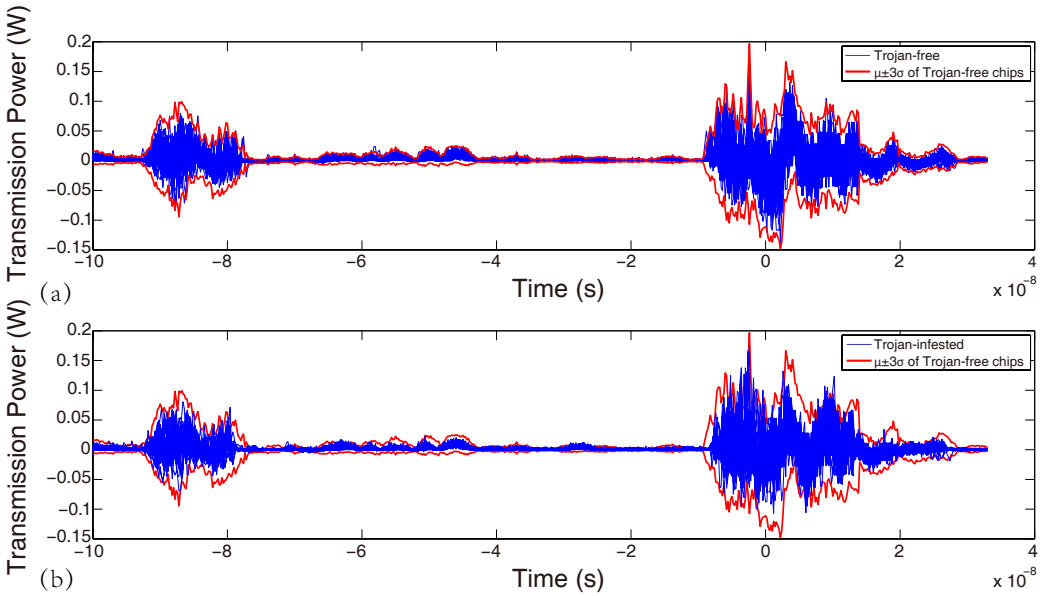


Fig. 10. Transmission power of (a) the 40 Trojan-free chips, and, (b) the 40 Trojan-infested chips, enclosed in the $\mu \pm 3\sigma$ envelope of the Trojan-free chips.

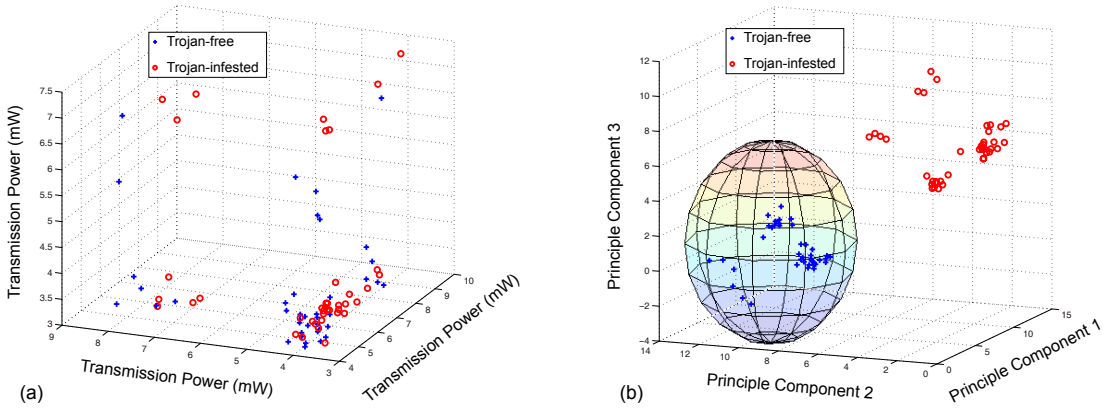


Fig. 11. Projection of hardware Trojan-free and hardware Trojan-infested circuits on a 3-dimensional space where each dimension corresponds to (a) the total transmission power for transmitting one cipher-text block, demonstrating that the populations cannot be separated in this space through simple upper/lower limits on each axis (i.e. through a hyper-cube); three out of the six blocks used in our experiment were randomly chosen and results are similar for any other subset of three measurements, and (b) one of the three top principal components yielded by performing PCA on the total transmission power for transmitting each of the six blocks for all the chips. The MVEE enclosing the hardware Trojan-free population, which can be used to classify a chip as hardware Trojan-free or hardware Trojan-infested, is also shown in this case.

[5] R. M. Rad, X. Wang, M. Tehranipoor, and J. Plusquellic, "Power supply signal calibration techniques for improving detection resolution to hardware Trojans," in *IEEE/ACM International Conference on Computer-Aided Design*, 2008, pp. 632–639.

[6] 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.

[7] C. Lamech, J. Aarestad, J. Plusquellic, R. Rad, and K. Agarwal, "REBEL and TDC: two embedded test structures for on-chip measurements of within-die path delay variations," in *IEEE/ACM the International Conference on Computer-Aided Design*, 2011, pp. 170–177.

[8] D. Du, S. Narasimhan, R.S. Chakraborty, and S. Bhunia, "Self-referencing: A scalable side-channel approach for hardware Trojan detection," in *Cryptographic Hardware and Embedded Systems*, pp. 173–187. 2010.

[9] K. Hu, A. Nowroz, S. Reda, and F. Koushanfar, "High-sensitivity hardware Trojan detection using multimodal characterization power mapping of integrated circuits using ac-based thermography," *IEEE/ACM Design, Automation and Test in Europe*, 2013.

[10] S. Narasimhan, D. Du, R.S. Chakraborty, S. Paul, F. Wolff, C. Papachristou, K. Roy, and S. Bhunia, "Multiple-parameter side-channel analysis: A non-invasive hardware trojan detection approach," in *IEEE International Symposium on Hardware-Oriented Security and Trust*, 2010, pp. 13–18.

[11] F. Koushanfar and A. Mirhoseini, "A unified framework for multimodal submodular integrated circuits trojan detection," *IEEE Transactions on Information Forensics and Security*, vol. 6, pp. 162 –174, 2011.

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

[13] I. T. Joliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[14] N. Moshtagh, "Minimum volume enclosing ellipsoid," in *http://www.seas.upenn.edu/ñima/papers/Mim_vol_ellipse.pdf, GRASP Lab, University of Pennsylvania*, 2005.