Hardware Trojan Detection with Linear Regression Based Gate-Level Characterization

Li Zhang and Chip-Hong Chang
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Abstract—**Due to outsourcing of IC fabrication, chip supply contamination is a clear and present danger, of which hardware Trojans (HTs) pose the greatest threat. This paper reviews the limitation of existing gate level characterization approaches to HT detection and presents a new detection method with a faster estimation of gate scaling factors by solving the normal equation of linear regression model. The HT-infected circuit can be distinguished from the genuine circuit without the need for a golden reference chip by their discrepancies in the bias parameter of the linear regression and a subset of the accurately estimated scaling factors. It has high detection sensitivity as long as the Trojan-to-circuit gate count ratio exceeds 0.4%.**

I. Introduction

To avoid the increasingly expensive capital cost of maintaining the semiconductor manufacturing facility, many semiconductor companies like Qualcomm, Broadcom and AMD have outsourced the fabrication of their chips to external contract foundries [1]. A rising concern of this merchant silicon business model is the trustworthiness of the manufactured chips. The IC supply chain can now be easily contaminated. There are many opportunities in the fabrication steps for an attacker to implant a malicious component in the design without changing the functionality. Such extraneous dormant function, known as hardware Trojan (HT), may act as a time bomb, leak secrets through side channel or allow the attacker to remotely take over the system containing the chip [2]. If these HT-infected chips are used in critical applications in financial, government or defense sectors, catastrophic damage may be inflicted.

Many post-fabrication methods have been proposed to detect potential HT infection. Destructive testing by micro-photography based reverse-engineering is too expensive and time consuming, and cannot be applied to all chips. Logic-test based approaches try to generate stochastic test patterns to activate the HT so that its effect can be detected at the circuit outputs. Such approaches are effective in activating ultra-small HTs but are feeble in triggering structurally and functionally complex HTs. HT may also be detected based on its manifested side-channel properties such as delay, transient current and leakage power. An anomaly in any of these properties indicates the existence of HT. Even so, their effect on the measured side-channel signals is easily masked by the large process variations (PVs) in today's nano-scale technologies.

From this perspective, gate-level characterization (GLC) approaches [3-6], in which PV is an innate attribute, outperform other side-channel based methods. Early HT detection methods assume the existence of a Trojan-free golden chip as reference for comparison with the IC under authentication (IUA). Such a golden chip is expensive to make and may not always be available. This requirement is avoided in GLC approaches by the linear extrapolation between the gate-level properties and measured side-channel signals. PV is an integral part of the gate property and is represented as a scaling factor to the nominal gate value in the system of linear equations (SLE). Existence of the HT will show up by the abnormal scaling factors in the SLE solution.

Existing GLC-based HT detection methods require expensive computations, series of statistical methods or a large number of measurements to ensure accuracy. A consistency based approach is proposed in [4], where a convex quadratic program is formulated with the objective function of minimizing the square root of measurement noise. The scaling factors are iteratively reweighted with the Gaussian kernel function. Existence of HT is verified by the large difference between the initial and final estimates of the scaling factor of each gate. Another consistency based approach [5] divides the IUA into multiple segments and performs the GLC for each segment with controlled input vectors. Overlapping gates across segments will have multiple estimated scaling factors. They will be consistent if all segments are Trojan-free and inconsistent if HT is present in one or more segments. The problem with this method is that it has limited sensitivity to HT that poses similar effects to all the affected segments. Solving convex quadratic problem iteratively is highly computational intensive and impractical for large IUA. A much faster linear program (LP) is formulated in [6] with an HT variable added into the SLE to indicate the presence or absence of HT. To improve the accuracy, thermal conditioning technique is employed to break the correlations among gates; the GLC process is repeated multiple times and maximum likelihood estimation is used to select the most likely values of scaling factors validated by statistical methods like re-sampling.

Post-silicon test time is expensive. Although precise GLC for each gate is helpful for HT detection and diagnosis, it is too slow for thousands of IUAs. In this paper, we propose a new approach to HT detection based on post-silicon leakage current measurement. In spite of some gate correlations, scaling factors of other circuit gates can be efficiently and accurately estimated by solving the normal equation in linear regression analysis. Based on the disparity of the bias term of linear regression and a subset of accurately estimated scaling factors, HT-infected chips can be easily distinguished from the genuine chip. As HT gates consume leakage power at all time, our method does not require the HT to be activated in order to detect it, making it possible to detect different types of HT, including those that are hard to be randomly triggered.

II. Preliminaries on Problem Formulation

A. Process Variation

A unique characteristic of deep sub-micron technologies is the intrinsic non-deterministic variations of process parameters. Such process variation (PV) is usually classified into two categories: inter-die and intra-die variations. Inter-die variations account for the variations arising between chips in the same or different wafers and its effect can be considered as being constant over a specific chip. In contrast, intra-die variations affect the devices in the same IC differently. Due to the normally distributed intra-die variations of device dimensions (like gate oxide thickness and effective channel length) and the exponential relation between these dimensions and the leakage current, the distribution of the leakage current variation is approximately lognormal [7]. Leakage current variation of a device can be deemed as an independent skewness that results in $I^r_{leak} = s \times I^n_{leak}$, where I^n_{leak} and I^r_{leak} are the nominal and real leakage currents for the device, respectively, and s is the scaling factor ascribed to the PV.

The scaling factor is usually assumed to be the same for different states of a device. These states typically correspond to different nominal leakage values, which are readily available from simulation models or other sources.

B. Gate-level Characterization of Leakage Current

By representing PV as a scaling factor of leakage current of each device, the total leakage current of an IC can be obtained by summing up the leaking current contributions from each gate on the chip. For every input vector, the state of each gate can be determined from the design information, e.g., the gate-level netlist. By summing up the scaled nominal leakage values of all gates and from the measured total leakage current for each input vector applied at the input pins, an SLE can be formulated. The i-th linear equation of the SLE is given by:

$$I^m_{leak}(v_i) = \sum_{j=1}^{N} s_j \times I^n_{leak,j}(v_i) \quad (1)$$

where N is the number of gates in the chip, $I^m_{leak}(v_i)$ and $I^n_{leak,j}(v_i)$ are respectively the measured leakage current of the chip and the nominal leakage current of gate j for the i-th input vector v_i, and s_j is the PV scaling factor for gate j.

As the measurements are usually taken at the external pins, they are highly susceptible to measurement errors due to variations in environmental conditions, thermal effects and noise. Taking the measurement errors into account, $I^m_{leak}(v_i)$ in (1) is the sum of the chip leakage current $I^c_{leak}(v_i)$ and the measurement error $e(v_j)$. The scaling factors are the variables to be estimated and can be calculated by solving the SLE with an objective function of minimizing the measurement errors.

C. Linear Regression with Multiple Variables

Let $x^{(i)} = (x^{(i)}_1, \ldots, x^{(i)}_N)$ be an input vector of N features and $y^{(i)}$ be the desired output. Assume that $D = \{D^{(1)} \ldots D^{(M)}\}$ is a set of M training data, where $D^{(i)} =< x^{(i)}, y^{(i)} >$ is the i^{th} data. A bias term $x^{(i)}_0 = 1$ is added to each input vector to learn the mapping from $x^{(i)}$ to $y^{(i)}$ for all $i = 1, \ldots, M$ by the linear regression model in (2). The optimal set of parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_N)$ can be obtained with the objective of minimizing the averaged sum of squared errors $J(\theta)$ in (3).

$$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}_1 + \ldots + \theta_N x^{(i)}_N = \theta^T \cdot x^{(i)} \quad (2)$$

$$J(\theta) = \frac{1}{M} \sum_{i=1}^{M} (y^{(i)} - \theta^T \cdot x^{(i)})^2 \quad (3)$$

The M input vectors $x^{(i)}$ (with $x^{(i)}_0 = 1$) form a matrix of dimension $M \times (N+1)$, denoted as X, while the M outputs $y^{(i)}$ form a vector denoted as y. Eq. (3) can be rewritten in terms of X and y in (4). The optimal set of linear regression parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_N)$ that minimizes $J(\theta)$ can be obtained when the partial derivative of $J(\theta)$ with respect to each parameter θ_j is 0, as expressed in (5).

$$J(\theta) = \frac{1}{M} (y - X \cdot \theta)^T \cdot (y - X \cdot \theta) \quad (4)$$

$$\nabla J(\theta) = -\frac{2}{M} X^T \cdot (y - X \cdot \theta) = \vec{0} \quad (5)$$

where $\vec{0}$ is a vector of $(N+1)$ 0's.

By rearranging the terms in (5), a normal equation (6) is derived for solving θ.

$$\theta = (X^T \cdot X)^{-1} \cdot X^T \cdot y \quad (6)$$

Even though $X^T \cdot X$ is non-invertible, the normal equation can be efficiently solved by using singular value decomposition to compute the pseudo-inverse of $X^T \cdot X$.

III. Proposed HT Detection Method

Our proposed HT detection method is based on the observed similarity between (1) and (2). The main idea is to map the scaling factor in the SLE $\sum_{j=1}^{N} s_j \times I^n_{leak,j}(v_i)$ to the parameters in the linear regression $\theta_1 x^{(i)}_1 + \ldots + \theta_N x^{(i)}_N$, while treating the measured leakage current $I^m_{leak}(v_i)$ with measurement error $e(v_i)$ as a predicted value of $h_\theta(v_i)$ with some prediction error. The remnant θ_0 can then be ascribed to the extraneous anomaly. By minimizing the measurement error, $I^c_{leak}(v_i)$ of an HT-free circuit can be construed as contributed entirely by the leakage of N circuit gates, and θ_0 is usually very small. Additional leakage contributed by the existence of HT to $I^c_{leak}(v_i)$ will thus be manifested by the abnormally large value of θ_0 or by the inconsistency in θ_0 values obtained from multiple runs with different test vector sets. The reason for the latter is that an HT is an extraneously introduced subcircuit and its leakage current will also vary with the input vector like any other circuit gates. To improve the HT detection sensitivity, instead of using the absolute value of θ_0,

we use the inconsistency of θ_0 under different test vector sets, measured by the variance of θ_0, to minimize the PV noise in the calculated parameters for each fabricated IC.

It turns out that the novel use of normal equation (6) gives a fast and good approximation of the scaling factors. If the input matrix X is full rank, the calculated θ_j (j = 1, ..., N) will match well with the actual scaling factor for each gate. However, due to gate correlations which make the coefficients of some gates collinear in X, the rank of X is usually smaller than N, resulting in only a portion of θ_j's being a good approximation of the actual scaling factor. The accurately estimated θ_j's, together with θ_0, are to be used as an indicator for the existence of HT, denoted by $Ind(\theta)$. By examining the mean and variance of each element in $Ind(\theta)$, an HT-infected circuit will be identified.

Our method consists of two phases, i.e., the HT indicator extraction phase and the HT detection phase. In phase I, the mean and variance of each element in $Ind(\theta)$ are to be collected. This phase can be performed by the design house using the simulation model of the circuit incorporating the PV, hence avoiding the need for an HT-free golden chip.

For a circuit of N gates, a set of $M = gN$ ($g > 1$) test vectors is randomly generated. The test vectors will produce an SLE of M linear equations in (1), from which an $M \times N$ matrix G of gate nominal leakage current and an M-element vector y of the measured leakage current can be constructed. The scaling factors can be calculated using the normal equation in (6) after augmenting the first column of G with an M-element column vector of all ones and forming the input matrix X.

The above steps for calculating θ are repeated K times, with a different random test vector set used at each time. The obtained K sets of θ are then post-processed to extract the HT indicators. For each θ_j (j = 0, ..., N), the K values below the 20th percentile and above the 80th percentile are curtailed to remove the influence of outliers. The remaining 60% of the K values are then used to calculate the mean and variance of θ_j. The parameter θ_j (j = 1, ..., N) is deemed to be accurately estimated if the fractional difference between the 20th percentile modulated mean of θ_j and the actual gate's scaling factor s_j falls within a small error margin ε (ε is set to be 0.01 or 1% in our case), as depicted in (7). The modulated mean and variance of θ_0 and the accurately estimated θ_j are then recorded to form the HT indicator $Ind(\theta)$.

$$\left|\theta_j - s_j\right| / s_j < \varepsilon \qquad (7)$$

In phase II, the K sets of test vectors used in phase I will be injected to the IUA and the total chip leakage current is measured for each set of test vectors. Multiple measurements can be performed for each input vector to reduce the random measurement errors. Based on the SLE formed under each test vector set, the matrix X and vector y can be constructed. Then, the normal equation (6) is used to solve for θ. Next, the same modulated mean and variance of each θ after removing the lower and upper 20th percentiles are calculated. The measured $Ind(\theta)$ of the IUA are compared with those recorded in phase I. If a large discrepancy exists, the IUA is considered to be HT-infected. Otherwise, the IUA is deemed to be HT-free. The flow of our HT detection method is depicted in Fig. 1.

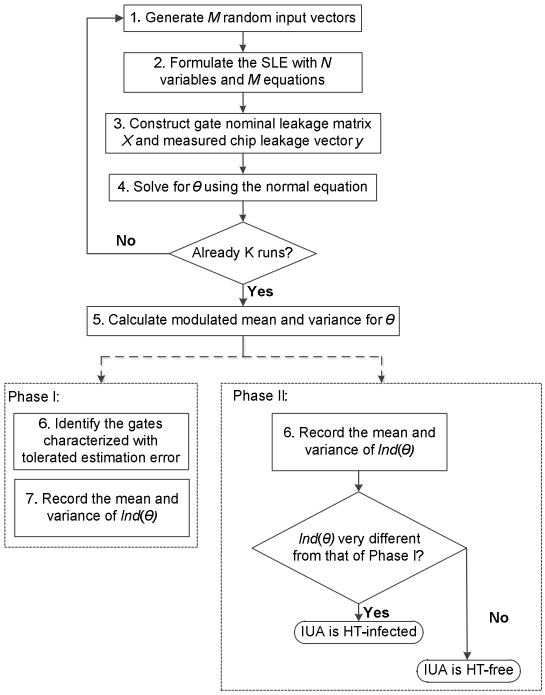


Fig. 1. Flow chart of the proposed HT detection method.

As an illustration, consider a circuit of seven NAND gates in Fig. 2. A total of K = 50 runs of tests are performed and each test vector set consists of 20 randomly generated vectors. For each test vector set, a 20×8 matrix X of nominal leakage current and a 20-element vector y of measured chip leakage current are constructed. By using the normal equation in (6), θ can be calculated. The modulated means and variances of θ for the seven gates of the genuine and HT-infected circuits are listed in Table I. The scaling factors s_j of each gate j obtained from the PV model used in the simulation are also listed. As Gates 3 and 7 always have the same inputs (i.e., they are correlated), their scaling factors, bold-printed in Table 1, cannot be correctly estimated. However, this does not influence the correct characterization of the other five gates. We will use the estimated scaling factors of these five gates' as well as θ_0 as the HT indicator, i.e., $Ind(\theta) = (\theta_0, \theta_1, \theta_2, \theta_4, \theta_5, \theta_6)$. It can be seen that there exists a large gap between $Ind(\theta)$ of the HT-free and HT-infected circuits, which provides a clear evidence to distinguish an HT-infected circuit from the genuine circuit.

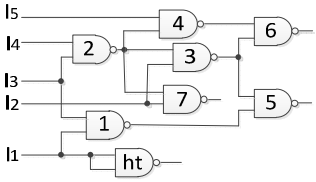


Fig. 2. The example circuit of 7 NAND gates with a 1-NAND HT.

IV. Experimental Evaluations

Our method is tested on circuits from the ISCAS'89 benchmark suit. Both the HT-free and HT-infected circuits are simulated. The nominal leakage values of the elementary gates are estimated by Cadence SoC Encounter with TSMC 0.18 µm CMOS standard cell library. It should be noted that the

successful HT detection rate of our method is not dependent on the choice of process technology since it is the relative difference among the leakage powers of the library cells instead of their absolute values that matters most [8]. A total PV of 12% is assumed in the simulations as in [9], where 20% of the PV is inter-die variation, 60% is spatially correlated intra-die variation and 20% is random uncorrelated intra-die variation. The measurement error is modeled using the triangular distribution with mean value of 1% as in [10]. Custom C++ codes were written to randomly generate the test vectors, deduce the final gate state for each vector, and construct the matrix X and y. The estimated parameters θ_j are obtained with Python codes which implements (6) and calculates the modulated mean and variance for $Ind(\theta)$.

TABLE I. The 20th percentile modulated mean and variance of the calculated θ with scaling factors s_j.

θ_j	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
s_j	NA	0.83	1.27	**0.92**	0.70	0.72	1.24	**1.50**
Mean(HT-free)	3E-4	0.83	1.27	1.21	0.70	0.72	1.24	1.21
Mean(HT)	–21	2.25	0.79	1.34	0.59	1.23	1.15	1.35
Var(HT-Free)	2E-4	9E-7	7E-7	6E-7	9E-7	1E-6	1E-6	6E-7
Var(HT)	41.35	0.18	0.09	0.12	0.15	0.22	0.19	0.12

The HT used is a 2-input, 1-output NAND gate well regarded as the most difficult case for HT detection [11]. Our experimental results show that the method can easily discriminate the HT-infected circuits as long as the Trojan-to-circuit ratio (TCR) in terms of gate count exceeds 0.4%. Due to the page constraint, two circuits C499 of 202 gates and C880a of 383 gates, representing the typical cases of HT detectable and HT indistinguishable respectively, are sampled for further analysis. Five instances of C499 circuits, each associated with a different PV, are tested. The simulation results for each of the genuine and HT-infected instances are shown in Table II.

TABLE II. 20th percentile modulated mean and variance of some elements of $Ind(\theta)$ from the five instances of C499 associated with different PV.

Instance	HT	θ_0 Mean (Var)	θ_{132} Mean (Var)	θ_{133} Mean (Var)
1	Free	–0.07 (0.39)	0.92 (6.6E-6)	0.89 (7.7E-6)
	Infected	–100.8 (7.17)	0.92 (9.3E-5)	0.89 (8.0E-5)
2	Free	–0.14 (0.56)	0.80 (8.3E-6)	0.81 (7.6E-6)
	Infected	–73.3 (5.27)	0.80 (6.8E-5)	0.81 (6.2E-5)
3	Free	0.04 (0.40)	0.77 (8.7E-6)	0.74 (9.5E-6))
	Infected	–140.6 (9.72)	0.77 (1.4E-4)	0.74 (1.1E-4)
4	Free	–0.26 (0.48)	0.96 (7.2E-6)	0.96 (9.8E-6)
	Infected	–74 (5.02)	0.96 (6.89E-5)	0.96 (6.0E-5)
5	Free	–0.05 (0.48)	0.74 (5.8E-6)	0.75 (9.0E-6)
	Infected	–106.2 (7.63)	0.74 (1.0E-4)	0.75 (8.7E-5)

From Table II, the variance of estimated θ is consistent among the five HT-free circuits. Hence, the variance of θ_0 and some correctly characterized gate's scaling factors can be used to detect the HT. The difference between the variances of the HT-free and HT-infected circuits is around 10~20 times. The gap is not obvious for C880a, which indicates that the HT-to-circuit gate count ratio has fallen below the HT detection sensitivity of our method. To confirm the case, two NAND gates are randomly inserted into C880a to double the ratio of HT-to-circuit gate count. The experimental results for a subset of $Ind(\theta)$ are shown in Table III. From Table III, the HT-infected circuit can now be easily distinguished from the HT-free circuit, as the TCR is now 0.52%, which is higher than the detection sensitivity of our method.

Table III. 20th percentile modulated mean and variance of some elementst of $Ind(\theta)$ from five instances of C880a associated with different PV.

Instance	HT	θ_0	θ_{191}	θ_{292}
1	Free	0.014 (3.5E-6)	1.24(4.0E-6)	1.503(2.73E-6)
	Infected	0.019(1.2E-3)	1.24(1.6E-3)	1.503(1.4E-3)
2	Free	0.016(1.7E-5)	1.635(5.4E-6)	1.069(1.96E-6)
	Infected	0.018(5.9E-4)	1.635(7.4E-4)	1.069(6.7E-4)
3	Free	0.013(8.2E-6)	1.369(4.8E-6)	1.02(2.8E-6)
	Infected	0.016(6.9E-4)	1.369(8.6E-4)	1.02(8.0E-4)
4	Free	0.014(1.5E-5)	1.132(5.48E-6)	0.589(2.1E-6)
	Infected	0.015(4.0E-4)	1.132(5.0E-4)	0.589(4.5E-4)
5	Free	0.014(1.8E-6)	0.758(4.6E-6)	1.36(2.7E-6)
	Infected	0.017(7.6E-4)	0.758(9.6E-4)	1.36(8.6E-4)

V. Conclusion

Our HT detection method efficiently solves the SLE formed by the GLC process. The scaling factors are estimated with linear regression. With the bias term of linear regression and a subset of accurately estimated scaling factors, our method is able to detect the presence of a wide range of HTs without the need for a golden reference nor the need to trigger the HT as long as the TCR exceeds 0.4%. The approach is an excellent complement to the logic testing approach which is effective in triggering ultra-small HTs. Without having to solve the GLC completely for every IUA at individual gate level, the time and computational cost required by our method is much lower than the conventional GLC approaches.

Acknowledgment

The authors would like to acknowledge the support from MOE-AcRF Tier2 grant MOE2013-T2-2-017.

References

[1] P. Clarke, "Spreadtrum, Dialog, MegaChips shine in fabless rankings," *EE Times*, April 12, 2012.

[2] R. Karri *et al.*, "Trustworthy hardware: identifying and classifying Hardware Trojans," *Computer,* vol. 43, no. 10, pp. 39-46, Oct. 2010.

[3] M. Potkonjak *et al.*, "Hardware Trojan horse detection using gate-level characterization," in *Proc. DAC*, July 2009, pp. 688-693.

[4] Y. Alkabani, and F. Koushanfar, "Consistency-based characterization for IC Trojan detection," in *Proc. ICCAD*, Nov. 2009, pp. 123-127.

[5] W. Sheng, and M. Potkonjak, "Scalable consistency-based hardware trojan detection and diagnosis," in *Proc. NSS*, Sept. 2011, pp. 176-183.

[6] S. Wei, and M. Potkonjak, "Scalable Hardware Trojan Diagnosis," *IEEE Trans. VLSI Syst.,* vol. 20, no. 6, pp. 1049-1057, June 2012.

[7] M. Potkonjak, "Non-invasive leakage power device characterization of integrated circuits using device grouping and compressive sensing," U.S. Patent US8286120 B2, Oct. 9, 2012.

[8] Y. Alkabani *et. al.*, "Input vector control for post-silicon leakage current minimization in the presence of manufacturing variability," in *Proc. DAC,* June 2008, pp. 606-609.

[9] D. Shamsi, P. Boufounos, and F. Koushanfar, "Noninvasive leakage power tomography of integrated circuits by compressive sensing," in *Proc. ISLPED*, Aug. 2008, pp. 341-346.

[10] S. Wei, S. Meguerdichian, and M. Potkonjak, "Malicious circuitry detection using thermal conditioning," *IEEE Tran. Inform. Forensics and Security,* vol. 6, no. 3, pp. 1136-1145, Sep, 2011.

[11] S. Wei *et. al.*, "Gate Characterization Using Singular Value Decomposition: Foundations and Applications," in *IEEE Trans. Inform.Forensics and Security*, vol. 7, no. 2, pp. 765-773, April 2012.