Scalable Hardware Trojan Diagnosis

Sheng Wei and Miodrag Potkonjak, *Member, IEEE*

***Abstract*—Hardware Trojans (HTs) pose a significant threat to the modern and pending integrated circuit (IC). Due to the diversity of HTs and intrinsic process variation (PV) in IC design, detecting and locating HTs is challenging. Several approaches have been proposed to address the problem, but they are either incapable of detecting various types of HTs or unable to handle very large circuits. We have developed a scalable HT detection and diagnosis approach that uses segmentation and gate level characterization (GLC). We ensure the detection of arbitrary malicious circuitry by measuring the overall leakage current for a set of different input vectors. In order to address the scalability issue, we employ a segmentation method that divides the large circuit into small sub-circuits using input vector selection. We develop a segment selection model in terms of properties of segments and their effects on GLC accuracy. The model parameters are calibrated by sampled data from the GLC process. Based on the selected segments we are able to detect and diagnose HTs by tracing gate level leakage power. We evaluate our approach on several ISCAS85/ISCAS89/ITC99 benchmarks. The simulation results show that our approach is capable of detecting and diagnosing HTs accurately on large circuits.**

***Index Terms*—Gate-level characterization (GLC), hardware Trojans (HTs), scalability, segmentation, thermal conditioning.**

I. Introduction

HARDWARE TROJANs (HTs) [2] are malicious hardware components embedded by adversaries in order to make the IC design malfunction or leak confidential information. Recently, HT attacks have drawn a great deal of attention in the hardware security community. The goal of HT diagnosis is to detect and locate the malicious HTs on the target circuit, so that they can be either masked or removed from the hardware. One of the ramifications of the current horizontal IC manufacturing model is that it is difficult to detect HTs during the manufacturing process, because anyone who realizes the circuit has complete access to the hardware and may conduct malicious modifications. Therefore, it is necessary and important to detect and diagnose HTs after manufacturing.

Among all the HT detection and diagnosis approaches, side channel analysis has been widely adopted because of its low instrumentation cost [3]–[7]. Side channel analysis detects HTs by observing the variations in IC manifestational properties such as delay, leakage power, and switching power. Since the presence of HTs would make at least one of the properties (e.g., leakage power) vary from its nominal specification, HTs are detectable by monitoring the delay/power characteristics of the circuit. In modern and pending technologies, process variation (PV) [8] is inevitable due to the nature of the IC manufacturing process. Although PV can be used to facilitate a variety of IC security applications, such as physically unclonable functions (PUFs) and true random number generators [9], it complicates the HT detection scheme, since any HT impacts can be easily explained as consequences of PV.

A number of recently published papers [9]–[14] proposed gate level characterization (GLC) approaches to capture HTs by tracing the side channels at gate level under the presence of PV. The existing approaches formulate a system of equations in terms of the gate level properties and the overall delay/power measurements. The presence of HTs results in inconsistent gate-level properties that are obtained from the solution of a system of equations. Nevertheless, there are two unsolved issues that may cause the failure of HT detection. First, in the system of equations, it usually happens that a large group of gates (variables) are correlated with each other in the sense that they have collinear coefficients due to the same gate type and same switching activities. Second, in modern IC design, especially with the fast development of deep submicron technologies, the transistor density keeps increasing, and it is common that a single IC has millions of gates. The huge number of gates results in large equations that require long running times.

We have developed a new approach for HT detection and diagnosis that employs a divide-and-conquer paradigm. The key idea is to divide the large circuit into small sub-circuits by using input vector control, so that the segmented circuits have desirable properties (e.g., small number of gates) for obtaining accurate HT detection results. We develop a segment selection model in terms of the properties of segments (e.g., controllability ratio and correlation ratio) and their impacts on the GLC accuracy. The model parameters are calculated based on the GLC accuracy results. We further introduce a test point insertion scheme to deal with possible large segments after the segmentation process. Next, we augment the capability of the segmentation approach using thermal conditioning-based GLC, which eliminates the collinear correlation problem by changing the thermal properties of the circuit and altering the coefficients in front of the scaling factors in the system of equations. Based on the segmentation approach we are able to obtain accurate GLC results and diagnose HTs correctly on large circuits by tracing gate level leakage power. The main contributions of this paper include the following:

- a segmentation technique that solves the scalability issue in IC characterization and HT diagnosis;

Manuscript received September 29, 2010; revised February 24, 2011; accepted April 02, 2011. Date of publication May 27, 2011; date of current version May 05, 2012. This work was supported in part by the NSF under Award CNS-0958369, Award CNS-1059435, and Award CCF-0926127. An earlier version of this paper was presented at the 2010 IEEE/ACM International Conference on Computer Aided Design (ICCAD 2010), pp. 483–486.

The authors are with the Computer Science Department, University of California, Los Angeles (UCLA), CA 90095 USA (e-mail: shengwei@cs.ucla.edu; miodrag@cs.ucla.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2011.2147341

- a segment selection model to select segments with desirable properties from a large IC;
- a test point insertion scheme to further split a large segment into smaller subparts;
- a systematic way of conducting HT detection and diagnosis using segmented GLC and constraint manipulation.

II. Related Work

Recently a number of approaches have been proposed for HT detection. Physical inspection [15] is a physical approach which scans the surface of an IC repeatedly in order to find the abnormal components. Many scanning techniques and visual inspection methods have been applied to scan the circuitry, but the downside of this approach is that it is too expensive to implement and thus not suitable for large scale HT detection. Functional testing [16] simulates the input vectors on the circuit and monitors the outputs to see if they match the expected patterns. This approach is easy to implement but it cannot detect the parametric Trojans that modify the characteristics of the original circuit but do not necessarily impact the functionality. Built-in-self-test (BIST) techniques [17] add additional circuitry to the target IC to monitor the abnormal signals, which are originally designed to detect manufacturing errors but are also applicable to HT diagnosis. BIST could detect the HT internally and thus give accurate results, but it requires extra circuitries for each specific design, which greatly increases the cost of the IC design.

In order to address the above shortcomings, gate level characterization [10], [13] was proposed for HT detection. The basic idea is to observe the circuit externally by characterizing the power/delay characteristics of all the gates. If HTs exist, their behaviors can be observed because they would add additional power consumption or delay to the original circuit. The basic assumption for using gate level tracing is that an accurate GLC is available for the circuit. However, due to the existence of PV, this goal is difficult to achieve, since it is hard to differentiate the variations of power caused by PV and caused by HTs.

GLC under the impact of PV is an attractive HT detection paradigm. The basic approaches that have been proposed [10], [12], [18], [19] characterize the physical properties of each gate by measuring the overall manifestational properties of the entire IC. A system of linear equations is obtained from multiple measurements based on the abstraction of PV scaling factors. A linear programming approach can be used to solve the system of linear equations and to obtain the characterization results. The current GLC approaches are able to obtain accurate results even if there are relatively large measurement errors, because they consider minimizing a norm of the measurement errors in the objective function so that the calculated measurement errors are close to the actual errors, and thus the characterized scaling factors are close to their actual values. However, what has not been considered in the current approaches is the scalability issue arising from the fact that there are a large number of gates (in the magnitude of millions) in modern IC designs. The resulting large systems of equations may easily exceed the computational limit of the LP solvers.

Furthermore, IC partitioning techniques have been proposed for scalable HT detection. Virginia Tech University research [7], [20] uses test generation techniques to isolate regions in an IC so that different levels of dynamic power increase can be detected between ICs with and without HTs. The essence of the approach is that IC partitioning is conducted in order to increase the likelihood of HT activation. Our segmentation-based HT detection approach is different from the Virginia Tech University research in the sense that we employ partitioning in order to obtain small segments of circuits and thus ensure scalability of HT detection. Also, there are numerous other differences including that the Virginia Tech University research targets only detection, but we conduct both detection and diagnosis of embedded malicious circuitry.

III. Preliminaries

We introduce the preliminaries and system models for our GLC and HTH detection process, including process variation, GLC, and thermal conditioning.

A. Process Variation

PV in IC manufacturing is the deviation of IC parameter values from nominal specifications, due to the nature of the manufacturing process [8], [21], [22]. PV is caused by the inability to precisely control the fabrication process at small-feature technologies [21]. For example, lithographic lens aberrations result in systematic errors on transistor sizes, and dopant density fluctuations impose random variations on design parameters. Also, PV impacts various levels of the IC design, including wafer-level, die-level, and wafer-die interaction [22]. The effect of PV plays a more and more important role with the rapid growth in CMOS scaling and performance enhancement.

There are two parameters that are greatly impacted by PV, namely threshold voltage $V_{\rm th}$ and effective channel length $L_{\rm eff}$. Due to the large variation in the two parameters, the parametric properties of an IC, such as propagation delay and leakage power, often deviate from the nominal design values. In this effect, the variation of $V_{\rm th}$ is especially important because it impacts both propagation delay and leakage power in a superlinear manner [21]. Borkar *et al.* [8] shows that there are wide variations of leakage power (up to 20×) and frequency (up to 30%) on a single wafer due to PV. It is commonly accepted that PV has shifted the design practice from the deterministic domain to the probabilistic and statistical domains [23], because the design process under PV has to deal with a range of possible parameter values. Therefore, it has been suggested that PV may wipe out most of the potential gains in design optimizations in the current and future technology generations [21]. We use the transistor model from Asenov *et al.* [24] and quad-tree spatial correlation model from Cline *et al.* [25] to generate instances of ICs that are subject to PV.

B. Gate Level Characterization

GLC is the process of characterizing each gate of an IC in terms of its physical properties (e.g., effective channel length) or manifestational properties (e.g., propagation delay or leakage/switching power). GLC has formed a natural basis

for side channel-based HT detection and diagnosis. We use the following gate-level power model [26]:

$$P_{\text{leakage}} = 2 \cdot n \cdot \mu \cdot C_{\text{ox}} \cdot \frac{W}{L} \cdot \left(\frac{kT}{q}\right)^2 \cdot V_{\text{dd}} \cdot e^{\frac{\sigma \cdot V_{\text{dd}} - V_{\text{th}}}{n \cdot (kT/q)}} \quad (1)$$

$$P_{\text{switching}} = \alpha \cdot C_{\text{ox}} \cdot W \cdot L \cdot V_{\text{dd}}^2. \quad (2)$$

Equation (1) is the subthreshold leakage power model [26], where L is effective channel length, V_{th} is threshold voltage, W is gate width, V_{dd} is supply voltage, n is substhreshold slope, μ is mobility, C_{ox} is oxide capacitance, ϕ_t is thermal voltage $\phi_t = kT/q$, and σ is drain induced barrier lowering (DIBL) factor. The gate-level switching power model [26] is specified in (2), where the switching power is dependent on oxide capacitance C_{ox}, gate width W, effective channel length L, switching probability α, and supply voltage V_{dd}.

From the power models, the impact of PV can be modeled as a PV scaling factor which indicates the increased/decreased power value due to PV. Therefore, a system of linear equations can be formulated in terms of the full-chip leakage power measurements and gate level power values:

$$\tilde{p_j} = e_{sj} + e_{rj} + \sum_{\forall gate\ i=1,\ldots,n} K_{ij} s_i \quad (3)$$

where $\tilde{p_j}$ is the full-chip leakage power for input vector j; s_i is the PV scaling factor of gate i; K_{ij} is the nominal leakage power for input vector j, which is dependent on the parameters in (1) and the input vector. Specifically, the value of K_{ij} can be found in a lookup table in [27]. e_{sj} and e_{rj} are systematic and random measurement errors, respectively.

C. Thermal Conditioning

One of the major issues in GLC is that a large number of gates are correlated with each other in the system of power measurement equations, because their coefficients (the nominal leakage power values) often have the same ratio [13]. This correlation issue is common in modern IC designs because gate replication is a common technique for improving speed. The correlation problem can be solved by thermal conditioning, where thermal control is applied on the correlated gates, and their nominal values are thus varied to break the collinear correlation. The idea is based on the observation that gate level leakage power increases exponentially with temperature. Specifically, another form of the leakage power model shows the relationship between leakage power and temperature [28]

$$P_{\text{leakage}} = A \cdot T^2 \cdot e^{\frac{\alpha v_{\text{dd}} + \beta}{T}} + B \cdot e^{\gamma v_{\text{dd}} + \delta} \quad (4)$$

where T is the temperature for a gate of interest, and A, B, α, β, and γ are empirical constants that can be found in [28]. The first term denotes the subthreshold leakage, which increases exponentially with the linear increase of temperature. The second term is gate leakage, which does not depend on temperature.

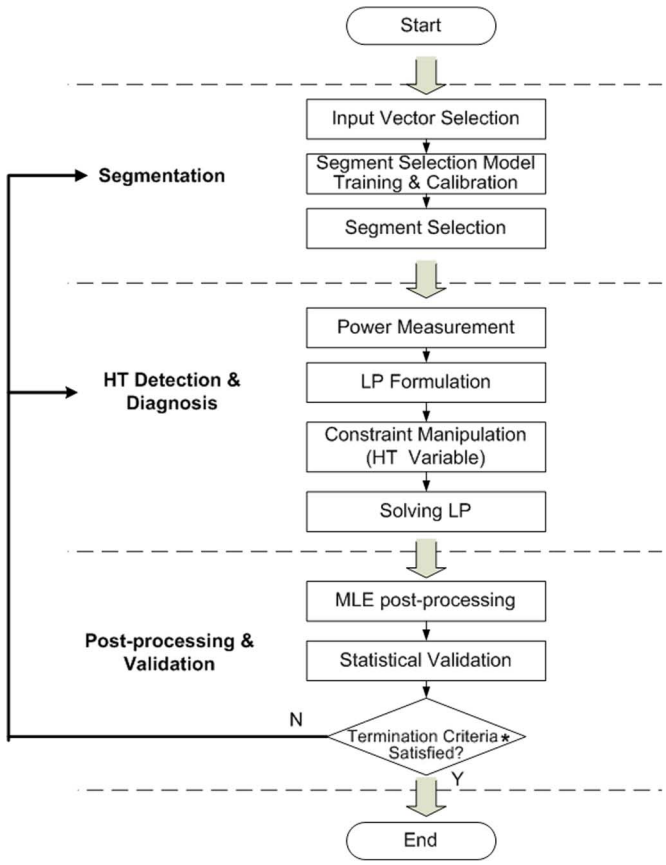

Fig. 1. Overall flow of our segmentation-based HT detection and diagnosis approach. We use a three-phase process including segmentation, HT detection and diagnosis, and post-processing.

IV. Overall Flow

In this section, we introduce the overall flow of our segmentation-based HT detection and diagnosis approach. The flow includes three phases, namely segmentation, HT detection and diagnosis, and post-processing, as shown in Fig. 1. In segmentation, our goal is to divide the circuit into small segments so that each individual segment has desirable properties (e.g., small number of gates) and all gates in the segment can be easily characterized by our GLC-based HT detection and diagnosis approach. We achieve this goal by selecting a number of segments as the training set and finding properties that impact the GLC accuracy of each segment. Based on the segment selection model, we select a set of most suitable segments that cover all the gates on the circuit. After segmentation, we begin the process of HT detection and diagnosis for each individual segment. In particular, we apply the set of input vectors and measure the total leakage power for each of them. For each measurement value, we formulate a linear equation by summing up the leakage power of each gate and considering the measurement errors. For HT detection, we further add a single HT variable in each of the linear equations that represents the presence of HTs. Next, we solve a system of linear equations using a linear program (LP) solver and obtain quantification for each PV scaling factor and the value of the HT variable in the case of HT detection. Finally, we repeat the process of GLC k times and conduct

post-processing based on the obtained results. We apply maximum likelihood estimation (MLE) that selects the most likely scaling factor values as our eventual results for GLC. Also, we adopt statistical methods (e.g., resampling) to validate our prediction results and repeat the previous segmentation and GLC steps if necessary. The entire GLC process terminates when the validated GLC accuracy is within a predefined threshold value.

V. Segment Selection Model

We solve the scalability issue in GLC as well as in the hardware security applications using segmentation techniques. The main idea is to partition the circuit into multiple small components and characterize each component of the circuit. In this section, we discuss our segment selection model for selecting segments that are amenable for accurate characterizations.

A. Problem Definition

Our goal in segment selection is to find segments that are likely to enable accurate GLC. Our definition of a segment is based on the controllability of a set of inputs over a set of gates. In particular, we employ the following two definitions.

Definition 1: Controllability. A set of inputs has control over a set of gates if the output signals of the gates can be controlled by varying the set of inputs while freezing the other inputs.

Definition 2: Segment. A segment is a sub-circuit that consists of gates and interconnects that can be controlled by a subset of inputs to the maximal possible extent.

We measure the GLC accuracy using relative characterization error, namely the mean error of the characterized gate level scaling factor compared to the real value

$$E_{\text{avg}} = \frac{1}{n_g} \sum_{i=1 \ldots n_g} \left| s_{calc_i} - s_{real_i} \right| / s_{real_i} \tag{5}$$

where E_{avg} represents the GLC error; n_g is the number of gates in the circuit; and s_{calc_i} and s_{real_i} are the calculated scaling factor of gate i and its real value, respectively.

In order to evaluate a segment in terms of its resulting GLC accuracy, we define the following properties of a segment by considering the key factors that may affect the GLC process, namely the number of equations and variables, the degree of correlation, and the level of imbalance of the gate level scaling factors.

- *Controllability Ratio*. We define the controllability ratio as the ratio between the number of unique equations and the number of variables that are available in a segment. The number of different equations of a segment can be calculated as the number of input patterns that can be obtained by varying the set of input vectors. The controllability ratio is an indicator of the controllability in the segment (n_{input} is the number of inputs, and n_g is the number of gates):

$$P_{\text{ctrl}} = \frac{2^{n_{\text{input}}}}{n_g}. \tag{6}$$

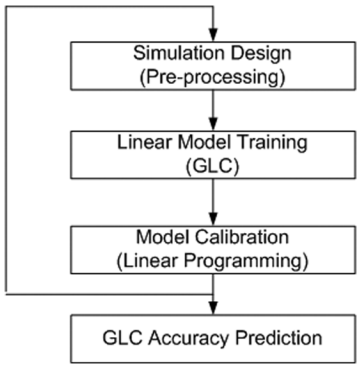

Fig. 2. Flow of segment selection model creation.

- *Correlation Ratio*. We define the correlation ratio as the ratio between the number of correlated gates in the system of linear equations and the total number of gates in a segment. The correlation ratio indicates the degree of correlation in the segment (n_{corr} is the number of correlated gates):

$$P_{\text{corr}} = \frac{n_{\text{corr}}}{n_g}. \tag{7}$$

- *Gini Coefficient*. We use the Gini coefficient [29] as a property that indicates the level of imbalance of the gate level scaling factors in a segment. The Gini coefficient can be calculated as the relative mean difference of a set of unordered data. In our case it is the difference between every possible pair of the scaling factors in the segment (μ is the average value of the scaling factor over all gates):

$$P_{\text{Gini}} = \frac{\sum_{i=1}^{n_g} \sum_{j=1}^{n_g} \left| s_{\text{real}_i} - s_{\text{real}_j} \right|}{2 n_g^2 \mu}. \tag{8}$$

- *Number of Gates*. We also consider the number of gates (n_g) in a segment as one of the properties. The intuition is that the size of the segment would impact the GLC accuracy due to the computational limit of the LP solver.

B. Model Creation

Our goal in segment selection modeling is to create a prediction model for GLC accuracy in terms of the properties of segments. Consequently, we use the model to predict the resulting GLC accuracy when selecting different segmentation strategies.

The basic flow of the model creation is shown in Fig. 2. We first select a set of segments from the target circuit so that a good and small enough set of training data can be found that helps make the predictors more accurate. Next, we conduct GLC on each selected segment and observe the GLC accuracy under the properties of each segment. After collecting the values of properties and GLC accuracy, we build up a regression model for the GLC accuracy prediction. We calibrate the model parameters by solving a system of equations using linear programming. The model characterization is a recursive process, in which we use different sets of segments obtained from pre-processing and keep calibrating the parameters until the variation of parameters between consecutive runs is smaller than a user specified

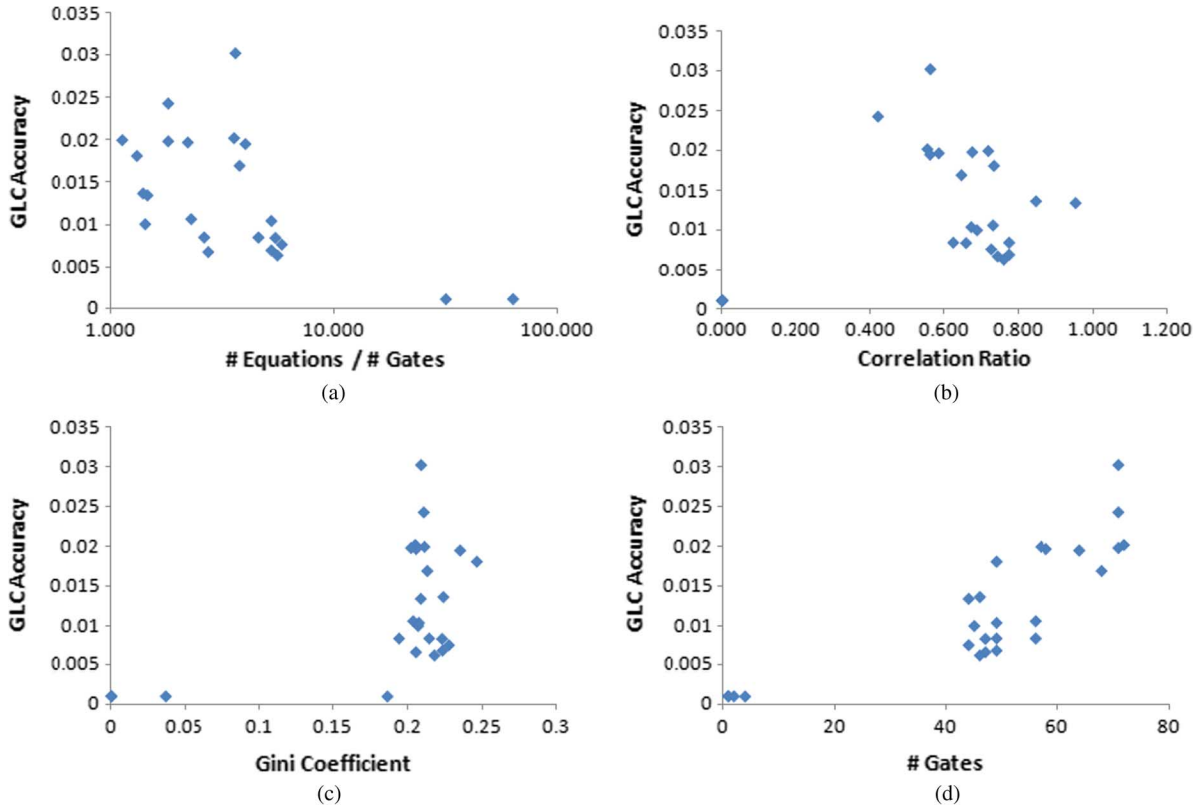

Fig. 3. Model training data for GLC accuracy and property values in ISCAS benchmark C432. (a) Controllability ratio. (b) Correlated ratio. (c) Gini coefficient. (d) Number of gates.

threshold. In the next phase, the prediction can be performed with respect to the other unknown segments.

1) Pre-Processing and Model Training: We conduct preprocessing by randomly selecting a small subset of the inputs. After that, we apply all the possible input vectors on the selected inputs and observe the gates controlled by the input vectors. We repeat the pre-processing and select many candidate segments for the model training. The property values of each segment can be calculated using (6) to (8). In model training, we conduct GLC on each segment and calculate the GLC accuracy. We are able to collect a group of training data in terms of GLC accuracy and the properties of the segments. Fig. 3 shows the data collected from the model training process on ISCAS benchmark C432. We observe that the relationship between the GLC accuracy and the property values match our intuition. For example, the GLC accuracy is improved when the controllability ratio increases, and when the correlation ratio decreases.

2) Model Characterization: Our objective in model characterization is to fit the training data into a closed form model, so that it can be used for GLC accuracy prediction as well as segment selection. In order to model the relationship between GLC accuracy and segment properties, we consider the following regression model:

$$y_i = \sum_{j=1}^{n} \alpha_j x_j + e_i \tag{9}$$

where y_i is the relative characterization error, when executing the ith sample of the training set. x_j is the value of the jth property in the training sample. The number of properties as well as the values of x_j are determined in the simulation design process. α_j is the prediction coefficient for the jth property. It is the key factor of our prediction model, which will be determined in the training process. e_i is the error when conducting the ith iteration of training. It represents the imprecision of the model when applied to predict the GLC accuracy in different segments. We assume that the errors follow a random distribution.

We regard (9) as a system of linear constraints in a linear program, in which x_j and y_i are constant values obtained from the training set, and α_j and e_i are unknown variables. Our goal is to minimize the possible errors in prediction, so we have the following objective function in the LP which minimizes the $l1$-norm of all the errors

$$\min \quad \sum_{i=1}^{m} |e_i|. \tag{10}$$

C. Segment Selection Algorithm

With the prediction model obtained in the training process, we are able to predict the GLC error according to the properties of a segment. We use the predicted GLC accuracy as a metric for segment selection. Our goal is to obtain a minimum GLC error using segment selection. The segment selection algorithm is a heuristic algorithm which keeps combining one-input segments in order to improve GLC accuracy. The combining operation stops and the algorithm starts with a new segment until there are no improvements in GLC accuracy according to the predic-

tion model. This is an iterative process where different starting segments are used in order to cover all the gates in the circuit. The segment selection algorithm is shown in Algorithm 1.

Algorithm 1: Segment selection algorithm

input: Input Set $PI = \{PI_i | 1 \leq i \leq n_i\}$

Controlled Gate Set
$G(PI_{sub}) = \{G_{sub} | G_{sub}\ is\ controlled\ by\ PI_{sub}\}$.

output: Selected Segment Set Seg.

1 **while** *not all gates covered in the circuit* **do**

2 Select starting PI_i close to the uncovered gates;

3 $PI_{sub} = \{PI_i\}$; // selected input set

4 **repeat**

5 $E_{cur} = E(PI_{sub})$; // current GLC accuracy

6 $Seg_{cur} = G(PI_{sub})$; // current segment

7 Add PI_j to PI_{sub}, where $PI_j \notin PI_{sub}$;

8 **until** $E_{cur} < E(PI_{sub})$;

9 Add Seg_{cur} to Seg;

10 Return Seg;

D. Dealing With Large Segments

After applying the aforementioned segment selection algorithm, we find that in the cases of large circuits, the number of gates in each segment is still very large. For example, in ISCAS89 benchmark S38584 there are around 875 gates in a single segment, and in ITC99 benchmark b17 the number is up to 1900. This is due to the fact that the depth of these large circuits is very high, and it is usually the case that the inputs have control over a large set of gates. In this case, the method of freezing a set of inputs cannot provide us with small segments.

We address this issue by further dividing the large number of gates in a single segment using IC partitioning techniques. We place additional controlling test points to the segments in the scan chain flip-flops, integrated with the regular scan chain that is used for manufacturing testing. In particular, we freeze the input that controls the maximum number of gates in the segment and replace it with an extra controlling test point that provides the maximum number of uncharacterized gates while keeping the segment size small. In order to find the place to insert the test point, we search the inputs of all gates in the segment and locate the input that maximizes the difference between the number of uncharacterized gates and the size of the segment

$$max \qquad (N_{\text{gates}} - N_{\text{seg}}) \tag{11}$$

where N_{gates} is the number of uncharacterized gates provided by the new test point, and N_{seg} is the segment size. The test point insertion is an iterative process, where we repeatedly add new test points until the resulting segment size is small enough to be processed by the LP solver.

VI. Scalable HT Detection and Diagnosis

We conduct HT detection and diagnosis using the segmentation approach. Since the HTs may be placed anywhere on the circuit, the goal of the diagnosis process is to detect the HTs and determine their locations if any exist. We achieve this goal by conducting thermally conditioned GLC on each segment of the circuit so that the scalability issue caused by large circuits can be addressed.

A. Technical Issues

HTs are difficult to detect and diagnose due to the following reasons. First, there are many types of HTs that have been made by adversaries. For example, there are functional HTs that affect the functionality of the IC, while other HTs are parametric that do not necessarily change the logic but rather modify the parametric properties of the chips (e.g., propagation delay or leakage power). To make things worse, there are HTs that are condition-based, which only trigger when certain conditions are satisfied (e.g., for a specific input vector). All of the above possible types make the detection of HTs non-trivial. Second, since the HTs can be placed anywhere on the chip and use any input signals, it is difficult to tell the exact locations even if we have detected their existence. Third, the adversaries try to insert small HTs into very large circuits so that the variations caused by HTs are hardly observable.

B. Objective and Flow

Our goal is to detect the existence of HTs (HT detection), and if there are any, to further identify their types, input pins, and locations (HT diagnosis). We use the side channel based method, namely GLC of the leakage power profiles, to identify HTs on the circuit. By using leakage power, we can ensure the detection of arbitrary malicious circuitry, because any HTs would cause systematic bias in the total leakage power, no matter they are activated or not. The main challenge we face is to make the small variation caused by HTs observable in large circuits. We solve this problem by using the aforementioned segmentation technique.

Our flow of HT detection and diagnosis is as follows. We first conduct preprocessing, in which we segment the large circuit into a few segments using Algorithm 1. Next, we apply thermal control on each segment and obtain leakage power measurements while varying the input signals. After obtaining the system of equations shown in (3), we manipulate the constraints and add one more variable (called the HT variable) into all the equations to indicate the variation caused by HTs. By solving the system of equations using the same method as in GLC, we obtain the characterization results for the HT variable and make conclusions about HTs based on its value. The detailed procedure of conducting HT detection and diagnosis is discussed in the next subsections.

C. HT Detection

HT detection indicates whether any HTs exist on the circuit or not. Our approach to HT detection is to employ a constraint manipulation paradigm based on GLC. The basic idea is that we first assume that HTs are present in the circuit. Since we do not know any information about the type, location, or input signals,

we just use a single HT variable to represent the existence of HTs in the LP formulation. Next, we solve the LP and check the value of the HT variable in the solution, which serves as an indicator for the existence of the HT. In particular, if the HT variable is close to 0, we conclude that no HTs are on the chip; otherwise, we confirm the presence of HTs and continue with the next step to diagnose their detailed characteristics.

Algorithm 2: HT diagnosis algorithm

input: 1. Circuit with HT for HT diagnosis;

2. λ, threshold value of HT variable that indicates the presence of HT.

output: L_{ht}, locations (inputs) of the HT on the circuit.

1 $L_{ht} = \emptyset$;

2 **for** *each segment of the circuit* **do**

3 **for** *each input* i *in the segment* **do**

4 Assume HT is embedded at input i;

5 Take measurements and formulate LP in the form of (13) with the HT variable var_{ht};

6 Solve the LP;

7 **if** $var_{ht} < \lambda >$ **then**

8 add i to L_{ht}

9 Return L_{ht};

The equations after constraint manipulation are the following, as modified from (3):

$$var_{ht} + K \cdot s = \tilde{p} + e \tag{12}$$

where var_{ht} is the HT variable we add as the indicator for HT.

D. HT Diagnosis

HT diagnosis is the process through which we infer the detailed information about the detected HTs, including their types, locations, and input signals. Our generic approach is exhaustive search, in which we first identify the type, location and input signals of the HT, and we verify our identifications by employing additional constraint manipulation based on (12), where we add a new HT variable according to our identifications

$$\mathrm{var}_{ht} + k_{ht} \cdot s_{ht} + K \cdot s = \tilde{p} + e \tag{13}$$

where $k_{ht} \cdot s_{ht}$ is the new item we added for HT diagnosis. k_{ht} is the leakage coefficient of the HT gate, which is dependent on the type, location, and input signals of the HT that we have identified. s_{ht} is the variable representing the PV scaling factor of the HT gate. After solving the LP, if the current identification is correct, we will obtain var_{ht} close to 0, and s_{ht} to be the estimated scaling factor for the HT. Otherwise var_{ht} is a large value which represents the discrepancy caused by an incorrect identification.

We show the HT diagnosis procedure in Algorithm 2. We examine each potential location using our generic GLC approach. As soon as we find that the HT value is much lower than those for other cases, we know that we have found the location of the HT. The key observation is that we conduct HT diagnosis on a per segment manner. Therefore, even if HTs may have multiple inputs (e.g., NAND gate), the running time is still relatively low because of the small segment size.

VII. Simulation Results

We evaluate our HT detection and diagnosis approach on the ISCAS85, ISCAS89, and ITC99 benchmarks. For each benchmark, we target two types of HT attacks, namely gate resizing and adding additional gates.

For the gate resizing attack, we employ segmentation-based GLC approach to characterize the size of each gate on a per segment manner. For the additional gates attack, we simulate two cases where the HTs do not exist or are embedded at random locations on the target circuit. We use as a HT a single inverter scaled to its smallest size and placed to a randomly selected location on the circuit. As discussed earlier, our HT detection scheme uses an extra HT variable as the indicator of HTs. We repeat the leakage power measurement 50 times so that we can have a large enough sample space to simulate the real measurement errors. By doing this, we expect to have a decision line which enables us to determine whether HTs exist or not. Table I shows the results in terms of the value of the HT variable we obtained for each set of measurements. We can see that for each benchmark the HT variable is always a large value (between 63 and 1349) when HTs exist, while it is a value close to 0 (between 0 and 15.8) when HTs do not exist. The results show that there is no overlapping, and we observe a large enough gap between the two situations. Therefore, a decision line can be obtained to determine whether there are any HTs embedded in the circuit.

After detecting any HTs on the target circuit, we start the HT diagnosis process to locate the HTs. We simulate the HT diagnosis process on the same benchmarks as in the HT detection process. In HT diagnosis, we use the HT variable as the indicator of a correct or incorrect identification. The identifications are based on the type, input signals, and the physical location of a HT. The results in Table II show that when the identification correctly identifies the actual location of a HT, the HT variable is small (from 0 to 2.5), while an incorrect identification always induces a large value (from 189 to 2615).

The number of guesses in our HT diagnosis is dependent on the number of gates being analyzed. If N is the total number of gates, the number of guesses in our HT diagnosis approach is $O(N)$ for 1-input HT, and $O(N^2)$ for 2-input HT, and in general $O(N^k)$ for k-input HT. For large circuits, the number of guesses can be very large. However, the essence of our approach is that this procedure is applied on individual segments instead of the complete design. Therefore, N is in the range of hundreds or less for each segment. Hence, we can detect even HTs with large numbers of inputs in practical amount of time. Even for our largest example, we finish in 10 hours. Also, since we are doing HT detection segment by segment, it is easy to parallelize the whole procedure.

TABLE I
HT DETECTION AND DIAGNOSIS ON ISCAS AND ITC BENCHMARKS

Benchmark	Gates	HT Detection		HT Diagnosis		
		HTs Exist	No HTs	Correct Identification	Incorrect Identification 1	Incorrect Identification 2
C17	6	226∼420	0	0	319	189
C432	160	608∼675	0∼11.0	0	709	765
C499	202	112∼297	0∼11.7	0.2	208	218
C880	383	63∼561	0∼7.2	0	234	237
C1355	546	229∼678	0∼3.0	0	508	448
C1908	603	592∼1282	0∼9.5	0.06	761	806
C2670	872	234∼1093	0∼0.09	0.03	335	395
C3540	1179	160∼1007	0∼0.3	0	674	583
S526	214	84∼724	0∼15.8	0	2603	2615
S832	292	230∼1272	0	0	2490	2494
S1423	490	571∼903	0∼0.001	0.005	257	795
S5378	1004	290∼1349	0	0.002	501	799
S9234	2027	752∼954	0∼0.005	0	423	954
S13207	7951	668∼1088	0	0	2038	421
S15850	9772	539∼857	0∼0.009	0	598	620
S35932	16065	503∼946	0∼0.01	0	693	326
S38584	19253	313	2.1	0	385	369
b17	32326	869	2.3	2.5	630	575
Smallest		63	0	0	208	189
Largest		1349	15.8	2.5	2603	2615

VIII. CONCLUSION

We have developed a scalable HT detection and diagnosis approach based on segmentation and gate level characterization. We ensure the detection of arbitrary malicious circuitry by measuring the overall leakage current, because any HTs must impact the overall leakage current even if they are not activated. By freezing a subpart of the input vectors and varying the others, we partition a large circuit into small components. We develop a segment selection model that predicts the GLC accuracy of the segments based on a learning data set and by linear programming. Next, a segment selection algorithm is used for selecting segments that result in small GLC error. We conduct GLC on each segment of the circuit and diagnose HTs by observing the variations on leakage power profiles. The simulation results on several ISCAS85/ISCAS89/ITC99 benchmarks show that HTs can be detected and diagnosed accurately on large circuits.

REFERENCES

[1] S. Wei and M. Potkonjak, "Scalable segmentation-based malicious circuitry detection and diagnosis," in *Proc. ICCAD*, pp. 483–486.

[2] M. Tehranipoor and F. Koushanfar, "A survey of hardware trojan taxonomy and detection," *IEEE Design Test Comput.*, vol. 27, no. 1, pp. 10–25, Jan. 2010.

[3] D. Agrawal, S. Baktir, D. Karakoyunlu, P. Rohatgi, and B. Sunar, "Trojan detection using IC fingerprinting," in *Proc. IEEE Symp. Security Privacy*, 2007, pp. 296–310.

[4] R. Rad, X. Wang, M. Tehranipoor, and J. Plusquellic, "Power supply signal calibration techniques for improving detection resolution to Hardware Trojans," in *Proc. ICCAD*, 2008, pp. 632–639.

[5] F. Wolff, C. Papachristou, S. Bhunia, and R. Chakraborty, "Towards trojan-free trusted ICs: Problem analysis and detection scheme," in *Proc. DATE*, 2008, pp. 1362–1365.

[6] J. Li and J. Lach, "At-Speed delay characterization for IC authentication and trojan horse detection," in *Proc. HOST*, 2008, pp. 8–14.

[7] M. Banga, M. Chandrasekar, L. Fang, and M. Hsiao, "Guided test generation for isolation and detection of embedded trojans in ICs," in *Proc. GLSVLSI*, 2008, pp. 363–366.

[8] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. DAC*, 2003, pp. 338–342.

[9] F. Koushanfar and M. Potkonjak, "CAD-based security, cryptography, and digital rights management," in *Proc. DAC*, 2007, pp. 268–269.

[10] M. Potkonjak, A. Nahapetian, M. Nelson, and T. Massey, "Hardware trojan horse detection using gate-level characterization," in *Proc. DAC*, 2009, pp. 688–693.

[11] M. Nelson, A. Nahapetian, F. Koushanfar, and M. Potkonjak, "SVD-based ghost circuitry detection," in *Proc. Inf. Hiding*, 2009, pp. 221–234.

[12] Y. Alkabani, F. Koushanfar, N. Kiyavash, and M. Potkonjak, "Trusted integrated circuits: A nondestructive hidden characteristics extraction approach," in *Proc. Inf. Hiding*, 2008, pp. 102–117.

[13] S. Wei, S. Meguerdichian, and M. Potkonjak, "Gate-level characterization: Foundations and hardware security applications," in *Proc. DAC*, 2010, pp. 222–227.

[14] S. Wei, S. Meguerdichian, and M. Potkonjak, "Malicious circuitry detection using thermal conditioning," *IEEE Trans. Inf. Forensics Security*, accepted for publication.

[15] J. Phang, S. Goh, A. Quah, C. Chua, L. Koh, S. Tan, and W. Chua, "Resolution and sensitivity enhancements of scanning optical microscopy techniques for integrated circuit failure analysis," in *Proc. IPFA*, 2009, pp. 11–18.

[16] A. Souza and M. Hsiao, "Error diagnosis of sequential circuits using region-based model," *J. Electron. Test.: Theory Appl.*, vol. 21, no. 2, pp. 115–126, 2005.

[17] G. Hetherington, T. Fryars, N. Tamarapalli, M. Kassab, A. Hassan, and J. Rajski, "Logic BIST for large industrial designs, real issues and case studies," in *Proc. ITC*, 1999, pp. 358–367.

[18] Y. Alkabani, T. Massey, F. Koushanfar, and M. Potkonjak, "Input vector control for post-silicon leakage current minimization in the presence of manufacturing variability," in *Proc. DAC*, 2008, pp. 606–609.

[19] F. Koushanfar, P. Boufounos, and D. Shamsi, "Post-silicon timing characterization by compressed sensing," in *Proc. ICCAD*, 2008, pp. 185–189.

[20] M. Banga and M. Hsiao, "A region based approach for the identification of hardware trojans," in *Proc. HOST*, 2008, pp. 40–47.

[21] S. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *IEEE Trans. Semicond. Manuf.*, vol. 21, no. 1, pp. 3–13, Jan. 2008.
[22] B. Stine, D. Boning, and J. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Trans. Semicond. Manuf.*, vol. 10, no. 1, pp. 24–41, Jan. 1997.
[23] S. Duvall, "Statistical circuit modeling and optimization," in *Proc. 5th Int. Workshop Statistical Metrol.*, 2000, pp. 56–63.
[24] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 um MOSFETs: A 3-D atomistic simulation study," *IEEE Trans. Electron Devices*, vol. 45, no. 12, pp. 2505–2513, Dec. 1998.
[25] B. Cline, K. Chopra, D. Blaauw, and Y. Cao, "Analysis and modeling of CD variation for statistical static timing," in *Proc. ICCAD*, 2006, pp. 60–66.
[26] D. Markovic, C. Wang, L. Alarcon, T. Liu, and J. Rabaey, "Ultralow-power design in near-threshold region," *Proc. IEEE*, vol. 98, no. 2, pp. 237–252, Feb. 2010.
[27] L. Yuan and G. Qu, "A combined gate replacement and input vector control approach for leakage current reduction," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 2, pp. 173–182, Feb. 2006.
[28] W. Liao, L. He, and K. Lepak, "Temperature-aware performance and power modeling," Univ. California, Los Angeles, Tech. Rep. UCLA Eng. 04-250, 2004.
[29] R. Keeney and H. Raiffa, *Decisions With Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge, MA: Cambridge Univ. Press, 1993.

Sheng Wei is currently pursuing the Ph.D. degree in computer science from the University of California, Los Angeles.

His research interests include computer-aided design of VLSI circuits, hardware security, and wireless networking.

Miodrag Potkonjak (M'02) received the Ph.D. degree in electrical engineering and computer science from University of California, Berkeley, in 1991.

He is a Professor with the Computer Science Department, University of California, Los Angeles. He created first watermarking, fingerprinting, and metering techniques for integrated circuits as well as first remote trusted sensing and trusted synthesis approaches, compilation using untrusted tools, and public physical unclonable functions.