Novel Techniques for High-Sensitivity Hardware Trojan Detection Using Thermal and Power Maps

Abdullah Nazma Nowroz, *Student Member, IEEE,* Kangqiao Hu, Farinaz Koushanfar, *Senior Member, IEEE,* and Sherief Reda, *Senior Member, IEEE*

***Abstract*—Hardware Trojans are malicious alterations or injections of unwanted circuitry to integrated circuits (ICs) by untrustworthy factories. They render great threat to the security of modern ICs by various unwanted activities such as bypassing or disabling the security fence of a system, leaking confidential information, deranging, or destroying the entire chip. Traditional testing strategies are becoming ineffective since these techniques suffer from decreased sensitivity toward small Trojans because of oversized chip and large amount of process variation present in nanometer technologies. The production volume along with decreased controllability and observability to complex ICs internals make it difficult to efficiently perform Trojan detection using typical structural tests like path latency and leakage power. In this paper, we propose a completely new post-silicon multimodal approach using runtime thermal and power maps for Trojan detection and localization. Utilizing the novel framework, we propose two different Trojan detection methods involving 2-D principal component analysis. First, supervised thresholding in case training data set is available and second, unsupervised clustering which require no prior characterization data of the chip. We introduce ℓ_1 regularization in the thermal to power inversion procedure which improves Trojan detection accuracy. To characterize ICs accurately, we perform our experiments in presence of realistic CMOS process variation. Our experimental evaluations reveal that our proposed methodology can detect very small Trojans with 3–4 orders of magnitude smaller power consumptions than the total power usage of the chip, while it scales very well because of the spatial view to ICs internals by the thermal mapping.**

***Index Terms*—2-D principal component analysis (2-DPCA), hardware Trojan detection, thermal and power mapping, unsupervised clustering.**

Manuscript received July 4, 2013; revised November 9, 2013 and March 17, 2014; accepted July 21, 2014. Date of current version November 18, 2014. This work was supported by the National Science Foundation under Grant 1115424 and Grant 1116858. An earlier version of this paper appeared at Design Automation and Test in Europe (DATE) 2013 [1]. This submission contains numerous novel materials, including: 1) detailed post-silicon multimodal thermal and power characterization framework with ℓ_1 regularization; 2) new unsupervised clustering method for Trojan detection; 3) impact of noise on thermal maps; 4) impact of increasing chip voltage; and 5) experimental results with new Trojan detection methods and new benchmark. This paper was recommended by Associate Editor Y. Xie.

A. N. Nowroz is with Intel Corporation, Austin, TX, USA (e-mail: mnowroz@gmail.com).

K. Hu is with Advanced Micro Devices, Inc., Austin, TX, USA (e-mail: kangqiaohu@gmail.com).

F. Koushanfar is with the Department of Electrical and Electronics Engineering, Rice University, Houston, TX 77005 USA (e-mail: farinaz@rice.edu).

S. Reda is with the School of Engineering, Brown University, Providence, RI 02912 USA (e-mail: sherief_reda@brown.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/TCAD.2014.2354293

I. Introduction

GLOBALIZATION of the semiconductor design and fabrication process due to the ever-increasing cost of manufacturing in small-scale CMOS technology has lead imminent threat to the security of integrated circuits (ICs). Besides foundry practices, modern IC design often use intellectual properties (IP) cores and electronic design automation (EDA) software tools, which are supplied by third party vendors. While the practice saves cost by utilizing the economy of scale, involvement of third-party entities exposes the chips by authentic designers to threats including hardware malware (Trojan) insertion, unlicensed IP handling, and IP piracy [2]–[4]. Since ICs form the core for the computing and communication systems used in contemporary personal, commercial, and government affairs, their exposure endangers the full systems built upon them. Therefore, developing noninvasive methods for screening and interrogating ICs for maintaining integrity in presence of unreliable third-party fabrication has become essential.

Hardware Trojans are implemented by unsought chip modifications by traitorously changing or tampering with the chips to provide opportunities for later exploits including controlling, monitoring, or spying the chip contents or secret keys [2], [4], [5]. Trojans can be very hard to detect, since they may be often inactive, only triggered as needed in target time intervals. Due to the increasing complexity of the contemporary chips and lack of controllability/observability to the chip internals post-silicon, the traditional structural and function tests are becoming ineffective in targeting Trojans. Invasive reverse engineering methods are slow, destructive, and expensive. Thus, devising noninvasive methods for examining the ICs and detecting Trojans has been recognized as a challenging research problem.

We devise a novel methodology for Trojan detection using multimodal post-silicon spatial thermal and power estimates. Chips can be thermally characterized using infrared emissions from the backside of silicon die, which then can be processed to get detailed spatial power maps [6]–[8]. These detailed thermal and power maps provide a much higher resolution Trojan detection method than previous methods where the total

current is measured and converted to gate-level. This detection procedure is easily scalable because the chip's spatial view in thermal mapping is not limited by the size of the chip and is only dependent on the thermal mapping resolution. The major contributions of this paper can be summarized as follows.

1) We propose a new direction for hardware Trojan detection and develop a multimodal characterization framework which includes spatial thermal maps and inverted power maps to detect and locate IC Trojans. Our detection framework involves acquiring post-silicon runtime thermal maps and applying residual inversion methods with l_1 regularization to obtain sparse spatial power maps which results in high sensitivity Trojan detection.
2) We employ 2-D principal component analysis (2DPCA) in order to tackle high dimensional thermal and power maps. Utilizing the 2DPCA framework, we present two different approach to Trojan detection, first is the supervised thresholding which needs training data set, and the second is unsupervised clustering method, which does not require any training data set.
3) To create realistic chips, we add 20%–40% process variations (PVs) to gate lengths, widths, and oxide thickness which can hide Trojans. To cover a wide range of variations, in our experiment we set five different PV levels with different standard variances which are obtained from realistic spatial variability models. We also add gaussian noise to our thermal maps to mimic real noise in infrared measurements.
4) We design virtual Trojans with local power density varying from 0.004 to 0.448 $\mu W/\mu m^2$ of total IC power consumption. To evaluate the accuracy of our Trojan localization method, we place the virtual Trojans in ten different locations in the chip.
5) We present an extensive set of simulation results with four different benchmarks with realistic chips and very small Trojan sizes. We show that our proposed methods are able to detect and locate Trojans with power consumption as small as 0.05 $\mu W/m^2$ very efficiently and accurately. We also evaluate the impact of thermal noise and chip voltage on the Trojan detection accuracy.

This paper provides a new dimension in identifying Trojans by using the spatial thermal and power information. The 2-D characteristics of the chip given by our method is not measurable by other noninvasive test methodologies such as standard delay, quiescent current in CMOS integrated IC (IDDQ), and transient power supply current (IDDT) tests. Note that our method has the ability to scale, because its resolution is independent of the chip size but dependent on the local Trojan power density (LTPD), which is defined as the power consumption of Trojan over the block size. Therefore, the valuable 2D information is used for detecting the Trojan and for identifying the problematic region as long as those regions are above the resolution of the method. Once those regions are identified, other localized testing methods, such as delay tests can be used for further enhancing the detection capabilities. Our method provides a new test modality which enhances and complements the resolution and performance of the existing Trojan detection methodologies that are based on noninvasive measurements from the chip.

The organization of this paper is as follows. Section II provides the necessary background. In Section III-B, we outline the thermal and power framework for our proposed Trojan detection procedure and in Section III-C, we describe our 2DPCA analysis. Section III-D describes various Trojan detection methods and Section IV describes our localization procedure. In Section V we discuss the impact of noise in thermal maps and measures to improve detection results. In Section VI, we present our experimental setup and results to demonstrate the effectiveness of our approach, and finally, Section VII summarizes our main results.

II. Background

A. Trojan Detection

Reports of instances of malware in military chips have triggered further research and investigations into the Trojan detection problem [3]. The utilized tests for Trojan detection include current-based methods, using static or dynamic currents [9]–[12], delay-based approaches [13]–[15], as well as simultaneous consideration of various current and delay testing methods [16]. In current-based approaches, both regional testing of current sums [9], [10], and translation of the currents to the gate-level [11], [12], [16] were pursued. While current-based methods can potentially provide a good characterization on smaller circuits, the presently available methods either need additional probes to the chip for regional current measurements [9], or necessitate formation and solving very large system of equations that are highly sensitive to noise and PV [11], [12], [16]. Delay-based detection methods have less components on each path and are easier to scale, but they suffer from the known problem of inadequacy of external test vectors for sensitizing all possible paths.

One of the work in this area [17] utilized the dynamic current (power) measurements by destructive testing of a few ICs from the design to build signatures. The assumption was that the fingerprint did not contain any malware. The existence of Trojan(s) in other chips were verified by noninvasively comparing against the signatures formed by destructive testing. Another path taken early for Trojan detection was to use verification and functional testing method. This approach simulates the inputs and then checks the corresponding outputs for the desired patterns [18], [19]. Functional testing suffers from the state-space explosion and lack of targeted verification output (since the Trojan behavior is not known in advance). Therefore, it's scope and effectiveness is rather limited.

The typical assumption for current- and delay-based Trojan detection approaches is that a golden model of the chip can be formed by post-layout simulations. The structural properties of the manufactured chips under investigation are then compared with this model. Such detection becomes more challenging for newer technology nodes which have surging random PV hard to separate from Trojan effect. What further complicates the problem is the large space of possibilities for Trojan exploit type and location.

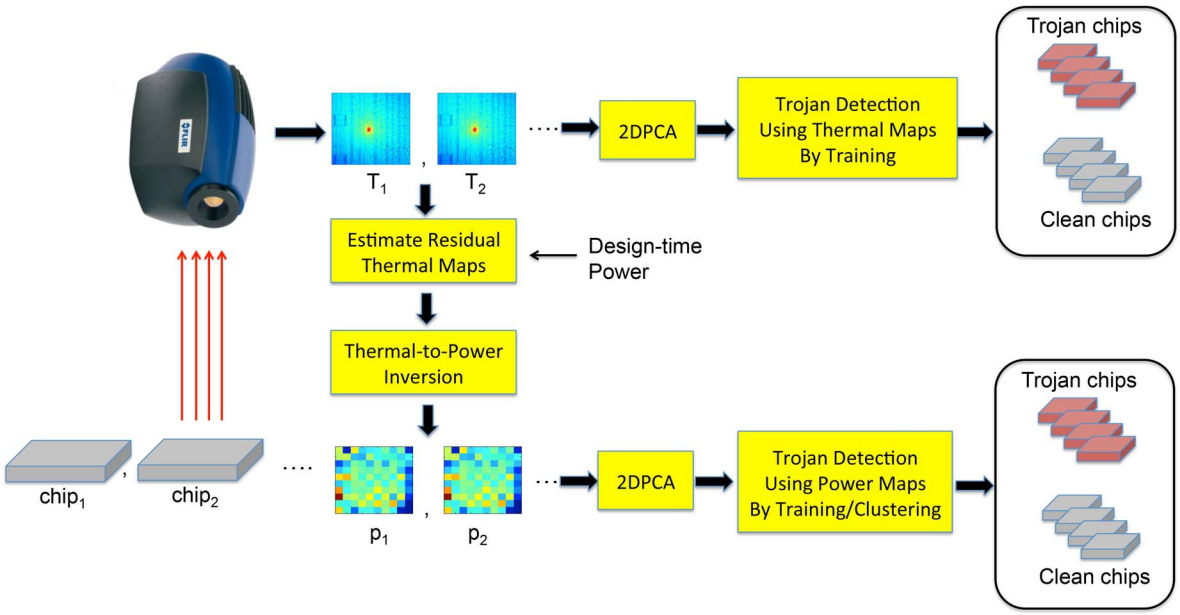

Fig. 1. Proposed Trojan detection framework using thermal and power maps.

An effective set of techniques pursued in this category is gate-level characterization [11], [12], [16], [20], which works well with both delay and current measurements. This method measures the chip's delay or current for a number of test vectors. Assuming that the currents (delays) linearly add up, a linear system is then constructed from the measurement set. Solving the system of linear equations translates the side-channel characteristics to smaller gate-level structural properties. While effective, this suite of techniques do not perform well for larger chips with more gates, higher accumulated measurement noise, and more sophisticated PV models. The existing gate-level characterization and Trojan detection techniques are evaluated on smaller benchmarks, where the performance of the method is the best. Note that, approaches based on regional testing of accumulated current which have a higher resolution only work for certain types of packaging and measurement probes [9], [10].

B. Post-Silicon Power Characterization

In post-silicon power mapping, a workload of a stable nature or a test pattern is applied to the chip under characterization, and the infrared emissions are captured from the back of the die using infrared imaging. The emissions are then inverted for obtaining the power maps. Due to the challenges in presilicon power modeling, a number of work have performed post-silicon power mapping to validate design choices and analyses [6], [7], [21]–[24]. While some of these works have been very successful in estimating the spatial power maps accurately, none of the previous work have utilized the spatial power maps to detect Trojans. We perform post-silicon power mapping on the high resolution thermal maps, but unlike previous work, we use the residual thermal maps instead of the actual thermal maps for Trojan detection. We exploit the fact that the residual thermal maps are sparse, and we add a regularization term in our quadratic programming to minimize the ℓ_1 norm of the power map. The details of our proposed power mapping is given in Section III-B2. Our proposed detection method utilizes these very high resolution thermal and power maps in order to detect IC Trojans which results into a very high sensitivity Trojan detection technique.

III. Proposed Trojan Detection Framework

Hardware Trojan detection is the process of detecting chips that are infected with unwanted Trojan circuitry and to verify the trustworthiness of the manufactured chips upon return to the clients. This new step requires defining a post-manufacturing step to validate the chip's conformance with the original specifications, which is called silicon design authentication. In this paper, we propose an entirely new multimodal framework for post-silicon Trojan detection using the thermal and power maps of the ICs running practical benchmarks.

Fig. 1 shows the framework of the proposed Trojan detection methods using post-silicon thermal and power characterization. In the beginning, workloads or test patterns are applied to the integrated chips and the runtime steady-state or averaged infrared thermal maps $\mathbf{T}_1, \mathbf{T}_2, \ldots$ are collected under realistic loading conditions. Using post-silicon thermal to power optimization framework, detailed residual power maps is constructed which has power dissipation in different hardware blocks. We propose a 2DPCA-based Trojan detection using the characterized thermal maps. More accurate Trojan detection technique using the reconstructed detailed power maps which requires the thermal to power inversion setup is presented. Depending on the availability of data from prior tests (training data) from chips that known to be benign, either of the two Trojan detection methods can be used on the residual power maps. The first requires a set of training chips to classify the Trojan infected chips and uses thresholding techniques. If no chips for training are available, then the second technique using unsupervised clustering can be applied. Note that we perform unsupervised clustering only with the inverted power maps, but not with the thermal maps, because

the natural clusters created by the features of the thermal maps are not properly distinguishable. We describe the details of our proposed framework components in the next sections.

A. Assumptions and Advantages

Like other hardware Trojan detection methods, one could detect the exploits as long as it lies within the assumptions and capabilities of the method. We make a few key assumptions in our research. First, during the test phase, the Trojans are assumed to be contributing to the current generated on the chip which may be of the form of dynamic or leakage current. Second, as long as the additional heat generated by the Trojan current is above the threshold of our measurement method, we are capable of detecting it. We do not put any constraints on how the Trojans are distributed, as long as their heat impact is above the resolution of the heat detector.

Here, we state advantages of using our heat detection method.

1) It is well known by the Trojan community that a strong attacker can always insert Trojans that are clandestine to all possible side-channel measurement methods [25]. The best a detector can do, to increase their chances for finding the Trojans that have an impact on the side-channels, is to measure as many side-channels as possible. The heat detection method, adds to the bag of available side-channel detection methods at the defender side. It's utility is orthogonal to the presently known techniques since it provides a 2-D view of the chip and a localization method which is not available by the previous methods based on timing or power measurement.
2) Our method has one key advantage since is much more scalable than the methods based on dynamic or static current alone. Simple current measurement methods constantly apply different test vectors and typically measure the cumulative current from one V_{dd} pin. As the chip size increases, the total current also increases, and the resolution of the method rapidly decreases. Since our method observes the heat from each of the 2-D spatial sub-regions of the chip, the heat measurement resolution is oblivious to the overall chip size. Therefore, our methodology is much more scalable than detection based on simple current analysis.
3) The heat detection method is also orthogonal to delay testing since it is well known that testing of all the sub paths is impossible because of controllability and output observability problems. Heat measurement does not rely on the observable timing difference in the output path to perform its test. Therefore, it brings in an advantage compared to Trojan detection based on timing.

B. Multimodal Thermal and Power Framework

We develop a new Trojan detection technique based on multimodal post-silicon spatial temperature and power characterization. The individual test modes are based on the chip temperature maps and power maps, respectively. The first mode in our framework corresponds to the high resolution thermal maps obtained using infrared imaging on post-silicon chips. Our purely thermal map-based Trojan detection method is able to detect and locate very small size Trojans. To further increase the Trojan detection resolution, we propose to invert the thermal maps to accurately estimate detailed spatial power maps, which corresponds to our second mode and can be used to perform power map-based Trojan detection. The more accurate power map-based Trojan detection enables detection of Trojans with as small power as 0.05%–0.2% of total power consumption of the chip. We describe our thermal mode and power mode as follows.

1) Thermal Mode: In the proposed procedure, infrared imaging is used to obtain thermal maps of post-silicon chips for Trojan detection. Modern integrated circuits use flip-chip packaging, where the die is flipped over and soldered to the package substrate. By removing the package heat spreader, one can obtain optical access to every device on the die through the silicon backside. There are two options to prevent any damage to the chips under test while removing the heat spreader. First option is that the heat removal system from a packaged chip is removed carefully without damaging the chip under test. After test, the chip can be repackaged and used as usual. Second option is to use a special socket board for testing before the packaging of the chip. The chips can be packaged for the first time after the Trojan detection tests are performed. Silicon is transparent in the infrared spectral region and this transparency allows the capturing of thermal infrared emissions using infrared imaging techniques [6]–[8], [23]. An infrared-transparent heat sink, for example, silicon window-based heat sink with mineral oil, has to be used to remove heat during operation of the IC [26].

For real chips, workloads of steady nature typically take around tens of seconds to reach the steady state. The thermal maps need to be captured after the chip temperature had reached steady state. Some workloads might not have a steady nature, in that case, the thermal maps can be captured for 30 s and averaged over time.

For the purpose of this paper, we first apply random vectors to the ICs and get the estimated power trace of each block by Primetime-PX. We then use HotSpot [27] thermal simulation tools to create the steady state thermal maps of various test bench circuits as described in Section VI-A1. We denote the steady-state thermal maps obtained using design-time simulations of the original authentic chip by $\mathbf{A}_1, \mathbf{A}_2, \ldots$ for each benchmark. We perform Monte Carlo simulations of the original chip at various PV corners to get power consumption under various PV scenarios. The thermal maps from chips under test by using infrared imaging is represented by $\mathbf{T}_1, \mathbf{T}_2, \ldots$ for each benchmark. It is possible to use the thermal maps for Trojan detection, but the sensitivity is less than the Trojan detection using power maps. If power mapping of the thermal maps is not available, these thermal maps can be used for Trojan detection as described in Section III-C. We use authentic thermal maps $\mathbf{A}_1, \mathbf{A}_2, \ldots$ as the training set and perform our Trojan detection methods of 2DPCA on the thermal maps under tests $\mathbf{T}_1, \mathbf{T}_2, \ldots$ for Trojan detection as described in Section III-C.

2) Power Mode: This section describes the power characterization of the chip. The power is obtained by inverting the thermal maps using quadratic optimization framework.

Procedure: Thermal to power inversion method
Input: Design time minimum power $\mathbf{p}_{min}$, Thermal maps under test $\mathbf{t}$, Thermal resistance matrix $\mathbf{R}$
Output: Residual power map $\mathbf{p}_r$

1) Find design time minimum thermal map, $\mathbf{t}_{min} = \mathbf{R}\mathbf{p}_{min}$;
2) Find residual thermal map, $\mathbf{t}_r = \mathbf{t} - \mathbf{t}_{min}$;
3) Solve quadratic programming: $\min \|\mathbf{R}\mathbf{p}_r - \mathbf{t}_r\|_2 + \|\mathbf{p}_r\|_1$, such that $\mathbf{p}_r \geq 0$.
4) Return solution of the quadratic programming $\mathbf{p}_r$

Fig. 2. Thermal to power inversion methodology.

The chip power and temperature are related by the heat equation, which can be discretized as follows by linear matrix formulation:

$$\mathbf{Rp + e = t} \tag{1}$$

where the 2-D thermal maps $\mathbf{T}_1, \mathbf{T}_2, \ldots$ is converted to vector representation $\mathbf{t}_1, \mathbf{t}_2, \ldots$ as required by the formulation in (1). These thermal maps give the measured temperatures at every pixel of the imaging system. The continuous power signal is represented by a vector $\mathbf{p}$ that gives the power density at a set of discrete die locations and the vector $\mathbf{e}$ denotes measurement noise in the infrared imaging system. The matrix $\mathbf{R}$ represents the thermal resistivities between different locations. The formulation of the matrix $\mathbf{R}$ is given in detail in [6] and [23]. For each specific chip, the matrix $\mathbf{R}$ can be estimated either by analytical methods, by simulation or experimentally on the real chip. We create matrix $\mathbf{R}$ by HotSpot simulation, by dividing the chip into 10×10 blocks and exciting each block at a time. The thermal map corresponding to one excited block represents one column in the matrix $\mathbf{R}$; this way we estimate matrix $\mathbf{R}$ for each chip column by column basis. The lower bound of the block size is limited by the resolution of infrared camera. The minimum resolution of a midwave infrared camera with appropriate optics is 5 μm. Detection accuracy increases as the block size decreases. There is a trade-off between the size of the blocks and computation time because as the block size decreases, the number of blocks increases, and Hotspot simulation time increases. Here, we make a trade-off between the resolution and accuracy based on our experiments.

Given a thermal map vector $\mathbf{t}$ and the matrix $\mathbf{R}$, the objective is to find the best power map vector $\mathbf{p}$ that minimizes the total squared error between the temperatures as computed from the estimated power $\mathbf{p}$ and the thermal measurements. For our case, we first subtract the thermal maps $\mathbf{t}_{\min}$ corresponding to minimum estimated design time power $\mathbf{p}_{\min}$, from the thermal maps $\mathbf{t}$ of chips under test, where $\mathbf{t}_{\min} = \mathbf{R}\mathbf{p}_{\min}$, and then invert the residual thermal maps, $\mathbf{t}_r = \mathbf{t} - \mathbf{t}_{\min}$, to get the residual power estimates $\mathbf{p}_r$. We want the estimates to be of the shape that only the blocks affected by Trojan have nonzero values while all other blocks remain zeros, which naturally leads us to finding sparse solution. Therefore, we add a regularization term in our quadratic programming to minimize the ℓ_1 norm of the power map, that is to minimize $||\mathbf{p}_r||_1$. The thermal to power inversion methodology is summarized in the algorithm given in Fig. 2. We apply our detection technique described in the following sections on the residual power maps.

We show an example of a thermal and power maps running workloads in advanced encryption standard (AES) cipher chip with 40% PV and 59.3 μW Trojan in Fig. 3, which are used for Trojan detection. Fig. 3(a) shows thermal map generated by HotSpot, Fig. 3(b) shows the residual thermal map after subtracting the design-time minimum thermal map. We divide the chip into 10×10 blocks and estimate the residual spatial power maps using optimization formulation. The chip dimension is 163×163 μm^2 and each block size is 265.7 μm^2. Fig. 3(c) shows residual power map estimated with previous thermal to power inversion method [1] and Fig. 3(d) shows the estimated residual power map using the proposed optimization formulation with ℓ_1 regularization. The Trojan location is shown in both the power maps. We can see that the power map become more sparse after using ℓ_1 regularization, which makes it easier to detect and locate the Trojan.

C. 2DPCA

Principal component analysis (PCA) is a classical feature extraction and data representation technique widely used in the areas of pattern recognition and computer vision. PCA is mathematically defined as an orthogonal linear transformation that translates the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. 2DPCA developed by Yang [28] is an image projection technique that makes use of the spatial correlation information to achieve better performance than conventional 1-D PCA [28]. The basic idea of 2DPCA is to project image $\mathbf{A}$, an $m \times n$ random matrix, onto a projection vector $\mathbf{x}$ by the following linear transformation:

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{2}$$

The discriminatory power of $\mathbf{x}$ is evaluated by the total scatter of the projected samples where the following criterion is adopted:

$$J(\mathbf{x}) = tr\,(\mathbf{S}_\mathbf{x}). \tag{3}$$

$\mathbf{S}_\mathbf{x}$ is the covariance matrix of the projected feature vectors of the training samples and $tr(\mathbf{S}_\mathbf{x})$ is the trace of $\mathbf{S}_\mathbf{x}$. The covariance matrix $\mathbf{S}_\mathbf{x}$ is given by the following equation:

$$\begin{aligned}\mathbf{S}_\mathbf{x} &= E\left[(\mathbf{y} - E\mathbf{y})\,(\mathbf{y} - E\mathbf{y})^T\right] \\ &= E\left[((\mathbf{A} - E\mathbf{A})\,\mathbf{x})\,((\mathbf{A} - E\mathbf{A})\,\mathbf{x})^T\right].\end{aligned} \tag{4}$$

So

$$tr(\mathbf{S}_\mathbf{x}) = \mathbf{x}^T E\left[(\mathbf{A} - E(\mathbf{A}))^T\,(\mathbf{A} - E(\mathbf{A}))\right]\mathbf{x} = \mathbf{x}^T\mathbf{G}_t\mathbf{x} \tag{5}$$

where $\mathbf{G}_t$ is the image covariance (scatter) matrix. Suppose there are totally M image samples for training, then

$$\mathbf{G}_t = \frac{1}{M}\sum_{j=1}^{M}\left(\mathbf{A}_j - \bar{\mathbf{A}}\right)^T\left(\mathbf{A}_j - \bar{\mathbf{A}}\right). \tag{6}$$

The optimal projection axes, $\mathbf{x}_{\text{opt},1}, \mathbf{x}_{\text{opt},2}, \ldots, \mathbf{x}_{\text{opt,d}}$, are the eigenvectors of $\mathbf{G}_t$ corresponding to the largest d eigenvalues.

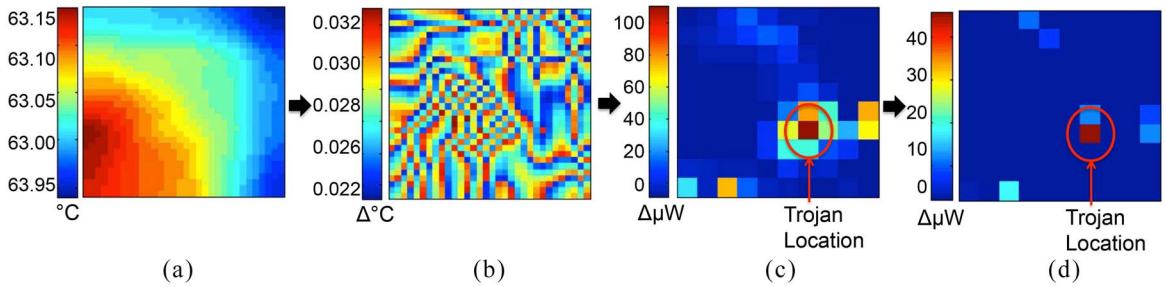

Fig. 3. AES cipher thermal map (°C) and estimated residual power map (μW). (a) Thermal map with Trojan. (b) Residual thermal map. (c) Residual power map without ℓ_1 regularization. (d) Residual power map with ℓ_1 regularization.

1) Feature Extraction and Identification: In our experiment, for purely thermal-based detection, 1000 thermal maps, $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{1000}$, of authentic chips are used to evaluate the optimal projection axes $\mathbf{x}_{opt,1}, \mathbf{x}_{opt,2}, \ldots, \mathbf{x}_{opt,d}$. Then the extracted thermal feature matrix $\mathbf{B}$ is defined by

$$\mathbf{B} = \left[\bar{\mathbf{A}}\mathbf{x}_{opt,1}, \bar{\mathbf{A}}\mathbf{x}_{opt,1}, \ldots, \bar{\mathbf{A}}\mathbf{x}_{opt,d}\right]. \tag{7}$$

For a given set of testing ICs, a feature matrix $\mathbf{B}_i$ is obtained for each IC after the transformation by 2DPCA. For power-based detection method, instead of using authentic thermal maps to build the feature matrix, the inverted power maps from authentic chips are used.

D. Trojan Detection Methods

We use the feature matrix $\mathbf{B}$ obtained by the 2DPCA analysis in Section III-C, for our Trojan detection. We perform Trojan detection in two different ways. First, if trusted data from known chips are available for training, a supervised thresholding method is used. Second, if no prior known data are available for training, unsupervised clustering technique is applied. We do not use any training chips for our unsupervised clustering technique. Both the techniques are described in following section.

1) Supervised Thresholding Method: The distance between the testing feature matrix $\mathbf{B}_i$ and the authentic feature matrix $\mathbf{B}$ is calculated by

$$d(\mathbf{B}, \mathbf{B}_i) = \|\mathbf{B}_i - \mathbf{B}\|_2 \tag{8}$$

where $\|\mathbf{B}_i - \mathbf{B}\|_2$ is the Euclidean distance between $\mathbf{B}_i$ and $\mathbf{B}$. If the distance is larger than a certain threshold, the testing IC is identified as Trojan inserted. The threshold is related to false positive rate (FPR) and obtained by applying the method to a set of authentic chips. For example, if we have 1000 authentic chips as training sets whose distances to golden chip are $d_1, \ldots, d_{1000}$, and we want to make the FPR within 1%, then the estimate of threshold is the value dividing $\{d_1, \ldots, d_{1000}\}$ into two sets, one of which has more than 990 chips and the other has less than 10 chips. The supervised thresholding method is applied to both thermal-based Trojan detection and power-based Trojan detection.

2) Unsupervised Clustering Method: Clustering is the most important unsupervised learning problem; it finds a natural grouping in a collection of unlabeled data and organizes objects into groups whose members are similar in some way [29]. As described in Section III-B2, we construct the residual power maps of the chips under test from the thermal maps. These detailed power maps with residual powers for each block have spatial groupings which can be used to distinguish chips under two different clusters, with and without Trojan. Since it is unsupervised, it means there is no learning step, and the algorithm does not need any prior knowledge, other than inputs which are the detailed power maps. This approach is suitable when we do not have a set of training chips in hand.

a) Appropriate feature selection: For clustering, it is very critical to choose an appropriate feature to be used for the partitioning. It influences the shape of the clusters as some elements may be close to one another according to one feature metric and farther away according to another. We have explored possible metrics to be used as a means for clustering our chips into two clusters of authentic chips and Trojan injected chips. We have the detailed residual power maps with $m \times n$ blocks of the chips under test, which we use for Trojan detection. To get a high resolution, we divide the chip layout into numerous blocks, which results in a high dimensional data set. One approach to cope with the problem of excessive dimensionality is to reduce the dimensionality by combining features. Some of the metrics or features that we have explored for clustering are maximum block power, variance among the block power, and spatial gradients among the block powers. The reason to use these features to distinguish between authentic and Trojan chips is that if there is a Trojan in the chip, then the maximum block power and variance among the block powers increases. Likewise, spatial gradients in the power maps also increase, which can help in detecting the Trojans. Another useful approach to tackle high dimensional data is to perform principal component analysis. We have explored the principal components derived by the 2DPCA analysis earlier and used the norm of the feature vectors found in the feature matrix, which yields the most accurate Trojan detection rate.

In Fig. 4, we plot the spatial gradients of the power maps in vertical and horizontal direction for PV of 20% and 40%. We can see that natural clustering patterns are prevalent in the power maps, and gradients is a suitable feature to distinguish between the authentic chips and Trojan chips when PV is 20%. We can observe that as the PV increases, the two clusters start to overlap, as a result some of the chips become unidentifiable which makes Trojan detection harder. This problem, which arises from the PV, can be overcome by using 2DPCA. Using the norm of the feature vectors found in the feature matrix $\mathbf{B}$ in

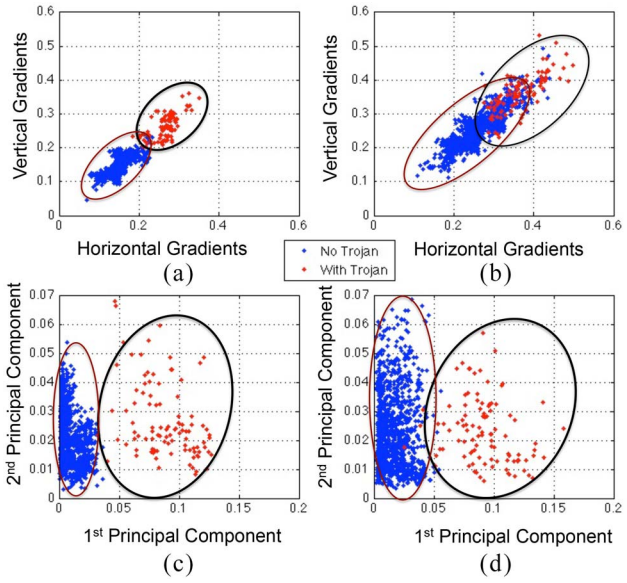

Fig. 4. Gradients of power maps with (a) PV 20%. (b) PV 40%. First and second component of feature matrix with (c) PV 20%. (d) PV 40%.

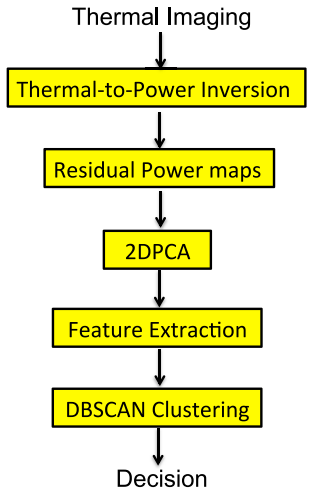

Fig. 5. Unsupervised clustering flow.

Section III-C, we observe that even with 40% PV, two clusters are properly distinguishable. We use the norm of the feature matrix obtained by doing 2DPCA on the residual power maps for our clustering feature. Our clustering process is shown step-by-step in Fig. 5. We propose to use spatial density-based clustering DBSCAN method to distinguish chips infected with hardware Trojan.

b) Density-based spatial clustering (DBSCAN): DBSCAN is a data clustering algorithm proposed by Ester *et al.* [30]. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. The advantage of DBSCAN is that unlike other popular clustering algorithms, such as k-means clustering, the accuracy of the clustering is not effected by the shape of the clusters. The spatial distribution of the clusters from authentic and Trojan chips can have any arbitrary shape making DBSCAN suitable for our case.

Procedure: DBSCAN based Trojan Detection
Input: Infrared-based residual power estimates for each block for chips under test, $\mathbf{P}_r$ as $m \times n_x \times n_y$ matrix, where m is the number of chips and $(n_x \times n_y)$ is the number of blocks
Output: Return Trojan infected chips

1) Perform 2DPCA on $\mathbf{P}_r$ to get feature matrix $\mathbf{B}$, which is $m \times n_x \times n_y$ matrix
2) For $i = 1, 2...m$:
   a. Find L_2 norm of $\mathbf{B_i}$, $\mathbf{P}_m(i) = ||\mathbf{B_i}||_2$;
3) Estimate ϵ and minPts from $\mathbf{P}_m$
4) Mark all points in $\mathbf{P}_m$ as unvisited
5) For each unvisited point p:
   a. Mark p as visited;
   b. Find ϵ-points, all neighbourhood points;
   c. **if** ϵ-points $<$ minPts: mark p as outlier
   d. **else if** p already in a cluster, add ϵ-points to the cluster
   e. **elseif** p not already in a cluster, start a new cluster and add ϵ-points to the cluster
6) Return the outliers as Trojan chip

Fig. 6. Unsupervised DBSCAN clustering for Trojan detection.

DBSCAN requires two parameters: the minimum number of points, min*Pts*, required to form a cluster, and *eps*, which is estimated from the data set under set by computing the geometric mean of the data. The ϵ-neighborhood is defined as the region that is covered with the given *eps*. The min*Pts* is the minimum number of members that a cluster can have. It starts with an arbitrary starting point that has not been visited. This point's ϵ-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. If a point is found to be a dense part of a cluster, its ϵ-neighborhood is also part of that cluster. Hence, all points that are found within the ϵ-neighborhood are added, as is their own ϵ-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise. DBSCAN algorithm for clustering is a well-known method, and there are various implements of the algorithm. We follow the DBSCAN routine formulated by Daszykowski *et al.* [31]. Our Trojan detection procedure is described in Fig. 6.

IV. Trojan Localization

The inherent low-pass filter of heat conduction function makes it hard to accurately locate the Trojan, since most of the high frequency components are lost in the thermal maps [6]. With the detailed spatial power characterization technique these frequency components are well recovered in the power maps. We use the estimated residual power maps to

locate the Trojans in the chip by finding the maximum power location in the Trojan detected chips

$$\arg\max_{i,j} \mathbf{p}_{r(i,j)} \tag{9}$$

where $\mathbf{p}_r$ is the estimated residential power map and (i, j) is the grid position index. For more generalized cases, such as ICs with multiTrojan, we detect local maxima points as the possible positions of the Trojans.

V. Impact of Thermal Imaging Noise on Trojan Detection

In this section, we discuss various noise sources that are present in the thermal imaging system and how it effects our proposed Trojan detection methods. We describe methods to mitigate the effect of noise to improve Trojan detection results.

The main sources of noise in the infrared imaging system are: 1) thermal noise; 2) digitization noise; 3) dark noise; and 4) flicker noise [32]. Thermal noise is caused by agitation of charge carriers and is present in all electronic devices. The analog-to-digital converter in the infrared camera causes the digitization noise. Dark noise is due to the random generation of electron-hole pairs in the quantum detectors which is usually present in photosensitive devices. Flicker noise is related to the trapping and detrapping fluctuations of charge carries at the transistor interfaces. The first three noise sources fall under the category of frequency-independent white noise which is the major source of noise in thermal imaging [33]. We mainly focus on mitigating the effect of white noise in our system in this paper.

By using a larger integration time (IT), the effect of white noise can be reduced dramatically. IT is defined as the time for which the thermal images for one single test are collected and averaged. As described in Section III-B1, the thermal maps in vector form is denoted by $\mathbf{t}$. Let $\mathbf{t_i(s)}$ denote the temperature of pixel i at time s as recorded by the thermal imaging system, and let I_p denote the integration period of the measurements. Then the temperature magnitude, $\mathbf{t_i}$, of pixel i is given by

$$\mathbf{t_i} = \frac{\mathbf{1}}{\mathbf{I_p}} \int_{\mathbf{0}}^{\mathbf{I_p}} \mathbf{t_i(s)ds}. \tag{10}$$

If there is no noise in the measurements, then we expect $\mathbf{t_i}$ to exhibit no stochastic behavior. However, noise in the measurements leads to a stochastic process where $\mathbf{t_i}$ is a random variable. Since white noise has a Gaussian distribution, by increasing the IT the standard deviation of the thermal signal which is responsible for the amplitude of noise can be reduced. From the central limit theorem, we know that the standard deviation of the average of a number of samples of random variable has $1/\sqrt{s_n}$ dependency on the number of samples, s_n. Thus, by increasing the IT, we can reduce white noise proportionally to the square root of IT [33]. As IT increases the noise is reduced, and it helps to improve Trojan detection rate. But if the workload does not reach steady state within the time when Trojan is activated, there will be a tradeoff between the IT and Trojan detection. In such a scenario, we can start with small IT, and increase the IT gradually. Moreover, we can select workloads that reach steady state in short amount of time for Trojan detection tests.

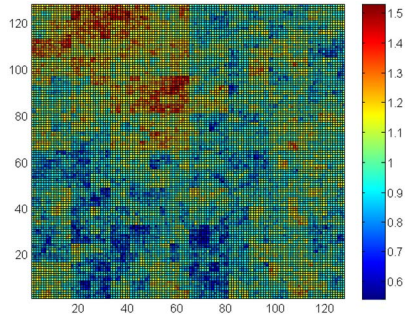

Fig. 7. PV profile. x-axis and y-axis are the two dimensions of the chip and the color of each grid represents the scaling factor of the gates in that grid.

VI. Experimental Setup and Results

To test our proposed Trojan detection methods, we provide sophisticated simulation results which mimic a realistic experiment setup with PV and then test four different benchmarks. We vary Trojan sizes and locations across the chips. We provide the experimental results of two approaches. First, the thermal map-based method which is more efficient in terms of computation time but less accurate; this does not require the thermal-to-power inversion procedure which increases detection time. Second, the detailed spatial power-based method is used which is very accurate and can detect and locate very small Trojan; this requires the residual thermal-to-power inversion with l_1 regularization procedure.

A. Experimental Setup

1) PV: To characterize real-world ICs accurately, we add 20%−40% PV to the gates' parameters. We use multilevel quad-tree approach to model the spatial within-die PV [34]. Higher levels of the quad-tree structure reflect the spatial correlations in larger scale while lower levels reflect the spatial correlations in smaller scale [34]. Fig. 7 demonstrates the PV profile generated by an eight-level quad-tree. The effect of PV on dynamic power is neglected in our experiment since it is insignificant compared to the effect of PV on leakage power. Since I_{sub} is the dominant component of leakage power, we assume that the leakage current is equal to sub-threshold current. We add PV to gates' length, gates' width and gates' oxide thickness as [35]. In our experiment we set five different PV levels with variation of 20%, 25%, 30%, 35%, and 40%, which introduces ±0.5% to ±3% variation to total power consumption.

2) IC Benchmarks: Four benchmarks from Opencores that are developed with hardware description language (HDL) are used in our analysis: 1) 128-bit AES cipher; 2) 32-bit microprocessor without interlocked pipeline stages (MIPS) Processor; 3) Reed–Solomon (RS) decoder; and 4) joint photographic experts group (JPEG). Table I gives the basic information of benchmarks including number of gates, core size, and total power consumption with standard voltage $1.1V$ at 1 GHz. We used design compiler synthesis tool from synopsys to map the benchmarks to Nangate 45 nm library and used Primetime-PX from synopsys to estimate the average power consumption during a certain period with random

TABLE I
TEST BENCHES

Test bench	No. of Gates	Core Size (μm^2)	Nominal Power (W)
AES	10610	163×163	0.0732
MIPS	8661	195×195	0.0494
RS Decoder	23224	394×394	0.12
JPEG Encoder	269970	1094×1094	1.4675

vectors. We used cadence SoC encounter register transfer language (RTL) compiler for floor planning, placing and routing, and Hotspot [27] for IC temperature simulation. The ICs are implemented with a constant core utilization of 70%.

3) Trojan Design and Insertion: We have designed Trojans modules with different power consumption. Our Trojans do not have any specific functional modules but certain power consumption triggered by the test patterns that are used to evaluate the minimum size of Trojan that can be detected. Despite the Trojan type, sequential or combinational, the LTPD is the only factor that impacts our detection results. We divide the IC area into 10×10 blocks and insert one Trojan per chip into the blank space within these blocks. The Trojan circuits are implemented using the same standard cells as the ICs with LTPD varying from 0.004 to 0.448 $\mu W/\mu m^2$, while the average chip power density is 1.26 $\mu W/\mu m^2$. We assume that the Trojan is inserted during the manufacturing stage through changing the mask, and the attackers tend to place the Trojan within a block instead of distributively due to the limited routing space. The impact of core utilization will be studied in the future work. For each benchmark at different PV levels, 10 000 chips with different sizes of Trojans inserted in different locations are generated. With different PV levels, different Trojan sizes, different Trojan locations 100 000 chips of each benchmark are generated for testing.

B. Results

We conduct and report the results of five experiments.

1) In the first experiment, we perform our supervised thresholding Trojan detection technique on high resolution thermal maps, and report results for four different benchmarks. We also analyze the effect of FPR on Trojan detection rate.
2) In the second experiment, we present results of two different Trojan detection methods using residual power maps. We assess our detection results with four different benchmarks and five different PVs.
3) We create Trojan infected chips with ten different Trojan locations. We compare our Trojan localization method under various benchmarks and PV.
4) The fourth experiment evaluates the effect of thermal noise in infrared imaging system on the detection results. We use different ITs and compare the accuracy of the Trojan detection.
5) In the fifth experiment, we increase the voltage of the chip from 1.1 to 1.2 V, and assess the effect of increasing voltage on the Trojan detection results.

1) Experiment 1: In the first experiment, we perform Trojan detection on high-resolution thermal maps. Based on the method proposed in Section III-C, we first calculate the optimal projection vectors for each benchmark. All the thermal maps are simulated by HotSpot in $2^n \times 2^n$ grids. n depends on the die size and the resolution of infrared camera, 5×5 μm^2. Thus, the thermal resolution of MIPS and AES is 32×32 grids, and RS Decoder and JPEG Encoder is 64×64 grids. The thermal maps with resolution $2^n \times 2^n$ have 2^n eigenvectors in total. The number of eigenvectors that are used for feature extraction is determined by the magnitude of corresponding eigenvalues. Here, we use benchmark AES as an example. We select eigenvectors corresponding to the first ten largest eigenvalues as the optimal projection axes. Then the average thermal map of 1000 authentic chips are used to extract the golden feature matrix **B** as shown in Fig. 8. For each chip under test, the distance of its feature matrix and the golden feature matrix is computed. Fig. 9 illustrates the distance distribution of authentic ICs and Trojan infected ICs. Fig. 9(a) is an experiment with 20% PV and Fig. 9(b) is the experiment with 40% PV and the same measurement error. From the figure, we can clearly see that as the magnitude of PV increases the histogram of distance begins to overlap, which makes it hard to distinguish the authentic chips from the Trojan infected chips. We have implemented experiments that vary the false positive and magnitude of PV.

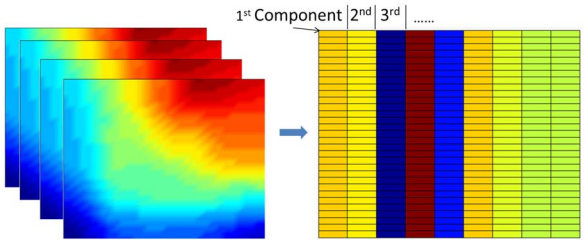

Fig. 8. Golden feature matrix extraction.

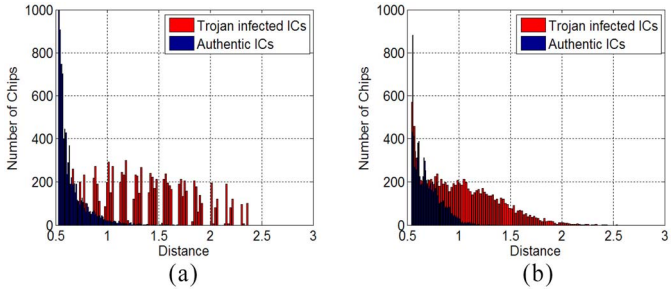

Fig. 9. Feature matrix distance between testing chip and golden chip AES under (a) 20% PV. (b) 40% PV.

As we mention in Section III-C, the testing IC instance is identified as an authentic chip or a Trojan infected chip by a certain threshold that is associated with detecting false positive. Based on the distance histogram (see Fig. 9) of training chips, we apply a kernel function to estimate the empirical probability distribution function (pdf) $f(d)$ for the authentic instances, where d denotes the distance from the golden feature matrix. Therefore, for a certain threshold d_{th}, the false positive is $\alpha = 1 - F(d_{th})$. By this, we fix the false positive to a certain value and observe how the false negative changes.

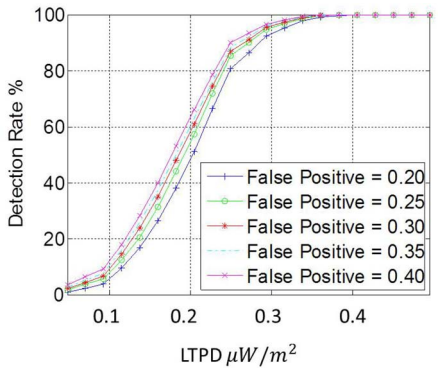

Fig. 10. Detection rate of AES under different false positives with fixed PV (0.2) and nominal voltage value (1.1 V).

TABLE II
TROJAN DETECTION RATE USING THERMAL MAPS

Benchmark	LTPD $\mu W/\mu m^2$\PV(%)	Detection Rate % 20	25	30	35	40
AES	0.148	8.0	2.0	2.2	2.7	3.0
	0.223	26.5	19.1	15.4	16.3	11.4
	0.297	86.5	67.2	54.4	43.5	31.2
	0.369	99.7	96.7	89.9	77.6	64.9
	0.443	100.0	100.0	100.0	96.5	90.8
MIPS	0.154	32.0	19.7	11.7	8.0	5.6
	0.231	60.6	39.0	23.3	16.5	10.8
	0.308	81.8	62.6	40.1	26.0	17.7
	0.384	91.1	77.8	58.2	39.4	27.6
	0.463	92.4	88.5	73.7	55.9	39.3
RS Decoder	0.062	5.4	5.0	2.7	2.3	2.3
	0.092	9.8	7.5	4.3	4.8	3.7
	0.122	21.2	13.4	9.1	7.3	4.8
	0.156	36.2	22.2	12.5	9.3	6.1
	0.218	57.0	38.1	20.9	12.8	10.1
JPEG	0.004	3.2	3.1	2.0	2.0	1.5
	0.015	8.0	7.8	7.5	8.0	3.3
	0.026	15.0	13.0	11.1	10.5	4.5
	0.037	25.8	21.2	20.6	15.5	10.3
	0.048	50.1	45.6	30.5	18.5	11.1

We define detection rate and FPR as follows:

$$\text{Detection Rate} = N_{TD}/N_T \tag{11}$$

$$\text{False Positive Rate} = N_{FD}/N_F. \tag{12}$$

Fig. 10 shows that as the false positive increases, the detection rate increases while the false negative decreases. The controllability of the threshold helps us to easily adjust the algorithm to trade off false alarm and detection rate according to different detection requirements.

a) Detection results under different PV level: The impact of PV is the most important factor that affects the performance of Trojan detection method. Table II shows that with the fixed FPR, as the magnitude of PV increases, the detection rate decreases. The detection rate decreases in the following order: AES, MIPS, and RS Decoder. The main difference among these three benchmarks are the total power and the core size. If we define power density, as $\rho = P/S_{\text{core}}$, where P is total power and S_{core} is the size of the core, we notice that ρ decreases in the same order as performance, which means $\rho_{\text{AES}} > \rho_{\text{MIPS}} > \rho_{\text{RS}}$. The chip with higher power density will generate more heat during the same period. Thus, a larger temperature gradient is formed, which makes the region with Trojan more prominent.

2) Experiment 2: In this experiment, we apply our supervised thresholding and unsupervised clustering technique proposed in Section III-D on detailed residual power maps. These high resolution power maps results in a very high sensitive Trojan detection. Overall, the power mapping approach has a much higher sensitivity than the thermal mapping approach. The reason for the dramatic improvement from thermal mapping to power mapping is the proper recovery of the high-frequency components of the power map due to the heat spreading information contained in **R** matrix. We plot detection rates for four benchmarks with tens of Trojan sizes and under five different PVs in Fig. 11, where N_{TD} is the total number of detected Trojan chip, N_T is the total number of Trojan chips, N_{FD} is the number of chips that is detected as Trojan and N_F is number of authentic chips.

We present detection results from our two different approaches, the supervised thresholding in Fig. 11(a) and unsupervised clustering method in Fig. 11(b). One can observe the trend that detection rate increases as the Trojan sizes increase, and decreases with PV as expected. From the table, we see that we are able to detect Trojan as small as 0.05 $\mu W/m^2$ with a detection rate above 50% and a false positive rate equal to 1%. The detection rate for each benchmark follows the same order as we have seen in the thermal map results in Experiment 1. The detection rate performance follows the same order as the power density of the chip, detection rate of AES > MIPS > RS_DEC > JPEG. Our results in Table III show that the detection rate decreases as the power density reduces. So the Trojan detection rate is not directly proportional to the area of the chip, rather it depends on the power density. For larger chip, it is possible to partition the chip into modules, and perform our method individually on each module. In that case the size of the chip should not affect our detection results.

For the supervised thresholding, we are able to fix the false positive rate to 1% for all cases of PV and Trojan size. Since the second method is unsupervised, there is no way to fix the false positive rate in that case. We add the corresponding false positive rate for the clustering method in Table II for comparison purpose. The false positive rate increases as the Trojan size decreases. For smaller Trojan sizes, we also see that false positive rate increases as the PV goes up. This is a limitation of the unsupervised method which is not the case for the supervised method. The advantage of the unsupervised is that we do not require any prior data set for the training purpose. Table III lists all the experimental results with residual power mapping.

There is a limitation in resolution like any other testing/detection method. For performing the thermal to power conversion, we divide the chip into 10×10 blocks power blocks, and estimate power for each block. This power resolution is limited by the thermal resolution of the infrared camera and the chip size. For our results, the power resolution can be increased more as long as it is less than the thermal resolution. The thermal resolution of MIPS and AES is 32×32 grids, and RS Decoder and JPEG Encoder is 64×64 grids. So if higher detection rate is required, then the chip can be divided into higher number of power blocks, increasing the

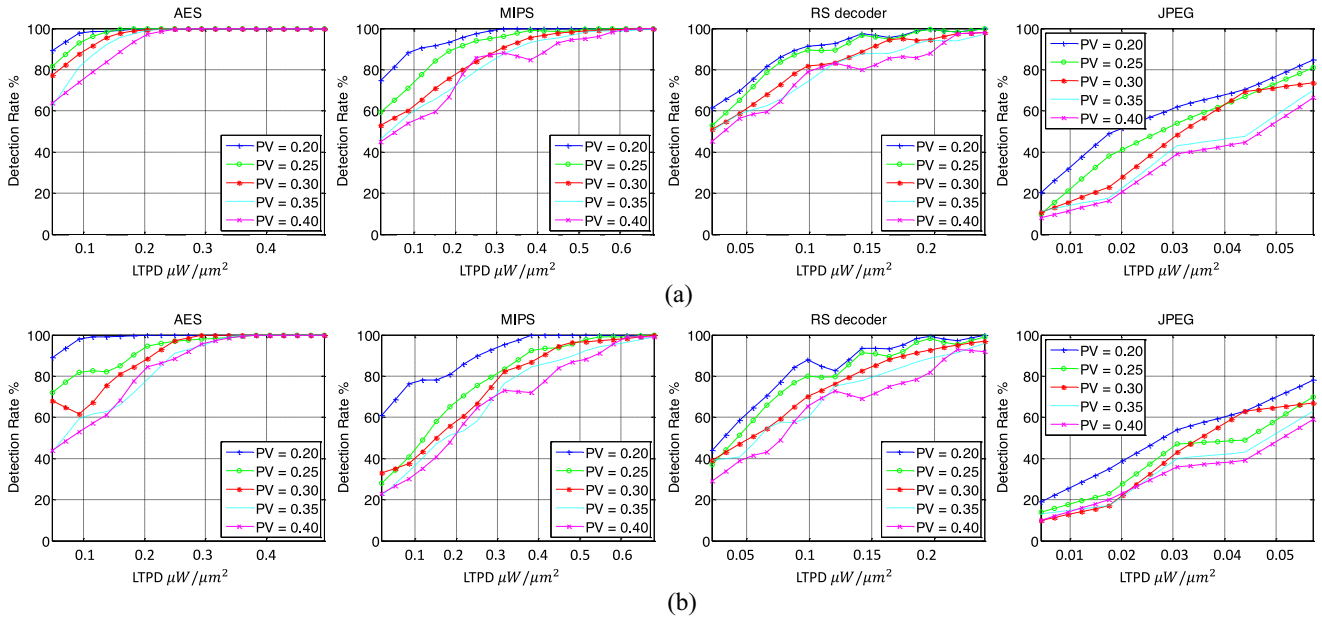

Fig. 11. Detection rates for four benchmarks using power maps (1.1 V), under different PV level. (a) Supervised thresholding technique with FPR 1%. (b) Unsupervised clustering technique with varying FPR shown in the Table III.

TABLE III
TROJAN DETECTION RESULTS USING POWER MAPS. LTPD: LOCAL TROJAN POWER DENSITY; FPR: FALSE POSITIVE RATE

Detection Method		Supervised Thresholding						Unsupervised Clustering									
		Detection Rate %					FPR %	Detection Rate %					FPR %				
Benchmark	LTPD $\mu W/\mu m^2$\PV(%)	20	25	30	35	40	20-40	20	25	30	35	40	20	25	30	35	40
AES	0.111	85	81	71	59	53	1	89	72	68	43	44	5	11	12	6	8
	0.148	98	93	86	80	66	1	99	83	61	61	54	5	10	4	6	7
	0.185	98	99	97	91	81	1	99	82	79	63	63	5	6	8	4	6
	0.223	100	100	99	99	97	1	100	94	87	76	84	5	4	5	4	5
	0.260	100	100	100	100	99	1	100	97	97	91	88	5	1	4	4	5
	0.297	100	100	100	100	100	1	100	98	100	95	96	5	1	4	2	5
	0.335	100	100	100	100	100	1	100	99	100	100	99	5	1	4	3	5
	0.369	100	100	100	100	100	1	100	100	100	100	100	5	1	4	3	5
	0.406	100	100	100	100	100	1	100	100	100	100	100	5	1	2	3	4
	0.443	100	100	100	100	100	1	100	100	100	100	100	5	1	2	3	4
MIPS	0.114	69	51	44	33	33	1	61	28	33	22	23	6	3	5	4	4
	0.154	85	65	51	49	49	1	78	42	38	35	31	6	3	4	5	4
	0.192	90	81	67	62	50	1	78	62	53	50	43	5	3	4	5	4
	0.231	96	90	76	72	77	1	89	74	64	55	63	5	2	4	4	4
	0.272	100	93	84	84	84	1	95	83	82	76	73	4	2	4	3	4
	0.308	100	99	93	90	80	1	100	93	87	85	72	5	3	4	3	3
	0.352	100	97	98	93	94	1	100	94	96	88	86	4	2	3	3	4
	0.384	100	100	97	98	94	1	100	99	97	94	89	4	2	3	3	4
	0.425	100	100	100	98	100	1	100	99	98	96	99	4	2	3	3	4
	0.463	100	100	100	99	100	1	100	100	100	99	99	4	2	3	3	4
RS Decoder	0.047	59	48	41	43	35	1	44	37	39	39	29	3	4	6	8	7
	0.062	70	66	57	53	47	1	60	53	48	41	40	3	4	4	6	7
	0.077	81	78	70	57	58	1	73	69	56	58	44	2	4	3	7	5
	0.092	91	91	80	71	74	1	89	80	69	57	64	3	3	3	4	5
	0.107	91	86	82	82	81	1	82	79	76	75	73	2	3	3	3	3
	0.122	98	97	88	86	76	1	94	92	83	78	69	1	2	3	4	4
	0.138	95	94	94	85	84	1	93	89	89	83	76	1	0	2	2	3
	0.156	100	100	94	94	82	1	100	99	92	88	79	1	0	2	2	3
	0.187	98	98	98	93	97	1	97	95	95	91	93	1	0	2	2	3
	0.218	100	100	97	97	98	1	100	99	97	96	92	1	0	1	2	2
JPEG	0.004	20	10	11	11	8	1	19	14	10	13	10	9	13	8	11	11
	0.015	49	38	23	18	16	1	35	23	17	17	20	4	8	7	11	11
	0.026	62	54	48	43	39	1	54	47	43	40	36	5	5	8	11	11
	0.037	70	67	69	48	45	1	63	49	63	43	39	3	3	6	7	7
	0.048	85	81	74	70	66	1	78	70	67	63	59	3	3	4	6	6

number of power blocks will improve the detection rate but simulation time will increase.

It is likely that adding other test modalities such as timing could improve the resolution but that would be orthogonal to the methodology presented in this paper. Having a fused detection method which integrates various modalities is a viable subject for future research. Another possible future direction is taking time-resolved measurements which requires a lot of additional work and is outside the scope of this paper.

3) Experiment 3: In this experiment, we present results of our Trojan localization method under various benchmarks and PVs. The sparsification process makes it very easy to localize

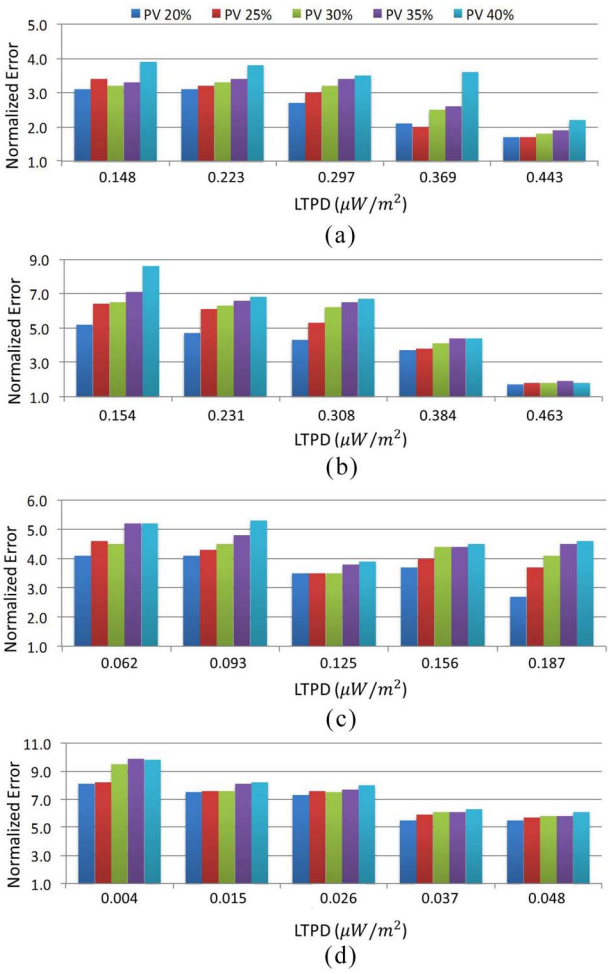

Fig. 12. Normalized localization error with five different PVs and four benchmarks. (a) AES. (b) MIPS. (c) RS Decoder. (d) JPEG.

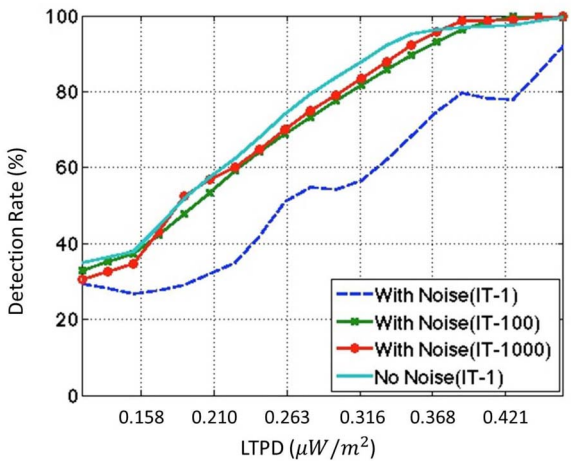

Fig. 13. Detection rate for MIPS with PV 30%.

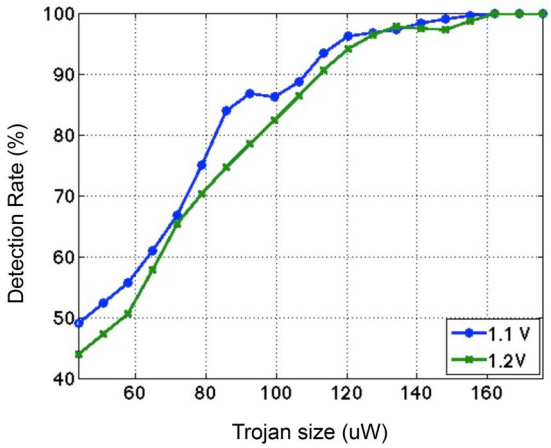

Fig. 14. Detection rate for MIPS with 1.1 and 1.2 V (PV 30%).

the Trojan. Once a chip is identified as an infected chip, we simply use the estimated residual power map to locate the Trojan by finding the maximum power location. We compute the localization error as the Euclidean distance between the estimated Trojan location and the real Trojan location. We normalize the localization error by dividing the error by the chip core dimension which is the length of one side of the core (e.g., for AES it is 163 μm). Fig. 12 shows the normalized localization error for four different benchmarks, AES, MIPS, RS decoder, and JPEG with five different PVs and five different Trojan sizes. We can see the Trojan localization error increases with increasing PV. Also, as the Trojan size decreases, it becomes more difficult to localize Trojan under the same PV.

4) Experiment 4: In this experiment, we add white Gaussian noise to our thermal maps to mimic a real infrared imaging system which has standard deviation of 10 mK. We select benchmark MIPS with PV of 30% for this experiment. We add gaussian noise to all our thermal maps, and then change our ITs. As discussed in Section V, the white noise in infrared imaging is inversely proportional to the square root of the IT. We then perform our residual power mapping as described in Section III-B on the integrated thermal maps. We apply the Trojan detection method with clustering as presented in Section III-D2 on the residual power maps. Fig. 13 shows the detection results. "With Noise (IT-1)" is the case where noise has been added to the thermal maps, and integrated over one frame, (IT-100) is integrated over 100 frames, (IT-1000) is integrated over 1000 frames. If the camera frame rate is 100 Hz, then 100 frames correspond to integration over 1 s, 1000 frames is integration over 10 s. "No Noise (IT-1)" stands for where no thermal noise has been added to the thermal maps. We see that as the IT increases, the detection rate with noisy thermal maps approaches the detection rate without noise. We conclude from this experiment that, the white noise that is present in the thermal maps can be compensated with increasing the ITs.

5) Experiment 5: We selected benchmark MIPS with PV 30% for this experiment and performed thresholding detection method. Fig. 14 shows the results for two voltage cases, 1.1 and 1.2 V. The plots show that there is not much difference since by increasing the voltage, the power of the Trojan and the power of variation caused by PV are both increased.

VII. Conclusion

In this paper, we have investigated the use of multimodal post-silicon spatial thermal and power maps in order to detect

and locate Trojans in modern ICs. We have developed two different Trojan detection methods, supervised thresholding and unsupervised clustering technique. Through an extensive set of benchmarks and experiments, we have demonstrated that using high resolution thermal maps increases the Trojan detection sensitivity. To improve the sensitivity further, we have inverted the residual thermal maps to detailed spatial power maps which are then utilized for Trojan detection. To exploit the sparsity of the residual thermal maps, we have added ℓ_1 regularization in our power mapping procedure which improves detection rate. These power maps can also reveal the Trojan location very accurately. Using proposed multimodal methods, we were able to detect Trojans which consume power as small as 0.05 $\mu W/m^2$. We have demonstrated that detection rate is directly proportional to the power density of the chip, and the Trojan size. To create realistic chips, we have added 20%–40% PVs, our results show Trojan detection is inversely proportional to PVs as it can hide Trojans. To mimic the thermal noise present in infrared imaging setup, we have added Gaussian noise to our thermal maps, and showed that the effect of infrared imaging noise on the detection rates can be mitigated by increasing the image IT. We have also compared detection results for different chip voltage settings.

We believe that our method has tremendous potential in solving many Trojan detection problems. For future work, we intend to develop a more general frame work that is applicable to a larger set of ICs and hardware Trojans. We will also explore the possibility in combining our method with other existing hardware Trojan detection methods, such as gate-level profiling technique, to achieve a detection resolution beyond the limitation of the camera. We are currently working on extending our method to pure leakage power-based detection to address the scenario where the Trojan will never be trigged during the test. The key problem we face in leakage power analysis is that the heat generated by pure leakage power is sometimes below the resolution of heat detector. This problem is solved by increasing the on-die temperature thus amplifying the leakage power. The detection results using amplified leakage power are comparable to those using dynamic power. Further, we are exploring a way to degrade the impact of PV using gate level profiling technique, which may result in increasing the detection rate and decreasing the false positive.

REFERENCES

[1] K. Hu, A. N. Nowroz, S. Reda, and F. Koushanfar, "High-sensitivity hardware Trojan detection using multimodal characterization," in *Proc. Conf. Des. Autom. Test Eur.*, Grenoble, France, 2013, pp. 1271–1276.

[2] M. Tehranipoor and F. Koushanfar, "A survey of hardware Trojans: Taxonomy and detection," *IEEE Des. Test Comput.*, vol. 27, no. 1, pp. 10–25, Jan./Feb. 2010.

[3] *Defense Science Board (DSB) Study on High Performance Microchip Supply* [Online]. Available: http://www.acq.osd.mil/dsb/reports/2005-02-hpms_report_final.pdf

[4] M. Tehranipoor *et al.*, "Trustworthy hardware: Trojan detection and design-for-trust challenges," *IEEE Comput. Mag.*, vol. 44, no. 7, pp. 66–74, Jul. 2011.

[5] J. Zheng and M. Potkonjak, "Securing netlist-level FPGA design through exploiting process variation and degradation," in *Proc. Int. Symp. Field-Program. Gate Arrays (FPGA)*, New York, NY, USA, 2012, pp. 129–139.

[6] R. Cochran, A. Nowroz, and S. Reda, "Post-silicon power characterization using thermal infrared emissions," in *Proc. Int. Symp. Low Power Electron. Des.*, Austin, TX, USA, 2010, pp. 331–336.

[7] H. Hamann, A. Weger, J. Lacey, Z. Hu, and P. Bose, "Hotspot-limited microprocessors: Direct temperature and power distribution measurements," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 56–65, Jan. 2007.

[8] F. J. Mesa-Martinez, E. Ardestani, and J. Renau, "Characterizing processor thermal behavior," in *Proc. 15th Edition Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, New York, NY, USA, 2010, pp. 193–204.

[9] R. Rad, J. Plusquellic, and M. Tehranipoor, "Sensitivity analysis to hardware Trojans using power supply transient signals," in *Proc. Int. Workshop Hardw.-Orient. Security Trust (HOST)*, Washington, DC, USA, 2008, pp. 3–7.

[10] R. Rad, X. Wang, M. Tehranipoor, and J. Plusquellic, "Power supply signal calibration techniques for improving detection resolution to hardware Trojans," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, Piscataway, NJ, USA, 2008, pp. 632–639.

[11] M. Potkonjak, A. Nahapetian, M. Nelson, and T. Massey, "Hardware Trojan horse detection using gate-level characterization," in *Proc. Des. Autom. Conf. (DAC)*, 2009, pp. 688–693.

[12] S. Wei and M. Potkonjak, "Scalable consistency-based hardware Trojan detection and diagnosis," in *Proc. Int. Conf. Netw. Syst. Security (NSS)*, 2011, pp. 176–183.

[13] Y. Jin and Y. Makris, "Hardware Trojan detection using path delay fingerprint," in *Proc. Int. Workshop Hardw.-Orient. Security Trust (HOST)*, 2008, pp. 51–57.

[14] J. Li and J. Lach, "At-speed delay characterization for IC authentication and Trojan horse detection," in *Proc. Int. Workshop Hardw.-Orient. Security Trust (HOST)*, 2008, pp. 8–14.

[15] S. Narasimhan, X. Wang, D. Du, R. Chakraborty, and S. Bhunia, "TeSR: A robust temporal self-referencing approach for hardware Trojan detection," in *Proc. Int. Symp. Hardw.-Orient. Security Trust (HOST)*, 2011, pp. 71–74.

[16] F. Koushanfar and A. Mirhoseini, "A unified framework for multimodal submodular integrated circuits Trojan detection," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 162–174, Mar. 2011.

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

[18] F. Wolff, C. Papachristou, S. Bhunia, and R. Chakraborty, "Towards Trojan-free trusted ICs: Problem analysis and detection scheme," in *Proc. Des. Autom. Test Eur. (DATE)*, Munich, Germany, 2008, pp. 1362–1365.

[19] M. Banga and M. Hsiao, "A region based approach for the identification of hardware Trojans," in *Proc. IEEE Int. Workshop Hardware-Oriented Security Trust (HOST)*, Anaheim, CA, USA, 2008, pp. 40–47.

[20] S. Wei, S. Meguerdichian, and M. Potkonjak, "Gate-level characterization: Foundations and hardware security applications," in *Proc. 47th ACM/IEEE Des. Autom. Conf.*, Anaheim, CA, USA, Jun. 2010, pp. 222–227.

[21] F. J. Mesa-Martinez, M. Brown, J. Nayfach-Battilana, and J. Renau, "Power model validation through thermal measurements," in *Proc. 34th Annu. Int. Symp. Comput. Archit.*, New York, NY, USA, 2007, pp. 302–311.

[22] Z. Qi *et al.*, "Temperature-to-power mapping," in *Proc. Int. Conf. Comput. Des. (ICCD)*, Amsterdam, The Netherlands, 2010, pp. 384–389.

[23] S. Reda, A. N. Nowroz, R. Cochran, and S. Angelevski, "Post-silicon power mapping techniques for integrated circuits," *Integr. VLSI J.*, vol. 46, no. 1, pp. 69–79, 2013.

[24] A. N. Nowroz, G. Woods, and S. Reda, "Power mapping of integrated circuits using AC-based thermography," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 21, no. 8, pp. 1398–1409, Aug. 2013.

[25] S. Wei, K. Li, F. Koushanfar, and M. Potkonjak, "Hardware Trojan horse benchmark via optimal creation and placement of malicious circuitry," in *Proc. 49th Annu. Design Autom. Conf. (DAC '12)*, San Francisco, CA, USA, pp. 90–95.

[26] S. Reda, R. Cochran, and A. N. Nowroz, "Improved thermal tracking for processors using hard and soft sensor allocation techniques," *IEEE Trans. Comput.*, vol. 60, no. 6, pp. 841–851, Jun. 2011.

[27] W. Huang *et al.*, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 5, pp. 501–513, May 2006.

[28] J. Yang, D. Zhang, A. Frangi, and J. Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, 2004.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2000.
[30] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, pp. 226–231.
[31] M. Daszykowski, B. Walczak, and D. L. Massart, "Looking for natural patterns in data. Part 1: Density based approach," *Chemometr. Intell. Lab. Syst.*, vol. 56, no. 2, pp. 83–92, May 2001.
[32] C. Hsieh, C. Wu, F. Jih, and T. Sun, "Focal-plane-arrays and CMOS readout techniques of infrared imaging systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 594–605, Aug. 1997.
[33] O. Breitenstein, W. Warta, and M. Langenkamp, *Lock-In Thermography: Basics and Use for Functional Diagnostics of Electronic Components*, 2nd ed. Berlin, Germany: Springer, 2010.
[34] A. Agarwal, "Statistical timing analysis using bounds and selective enumeration," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 22, no. 9, pp. 1243–1261, Sep. 2003.
[35] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations," in *Proc. Int. Symp. Low Power Electron. Des.*, 2002, pp. 64–67.

Abdullah Nazma Nowroz (S'10) received the degree (*summa cum laude*) in electrical and electronics engineering from Boston University, Boston, MA, USA, in 2006, the master's degree from the University of Southern California, Los Angeles, CA, USA, and the Ph.D. degree from Brown University, Providence, RI, USA, both in electrical engineering, in 2008 and 2013, respectively.

She is currently with Intel Corporation, Austin, TX, USA. Her current research interests include thermal and power modeling of integrated circuits and hardware security.

Ms. Nowroz is a member of the IEEE, ACM, Tau Beta Pi, Eta Kappa Nu, and National Society of Collegiate Scholars (NSCS).

Kangqiao Hu received the B.S. degree in electrical and electronics engineering from Tsinghua University, Beijing, China, and the M.Eng. degree from Rice University, Houston, TX, USA.

He is currently an Application Specified Integrated Circuit Design Engineer with Advanced Micro Devices, Inc., Sunnyvale, CA, USA.

Farinaz Koushanfar (S'99-M'06-SM'13) received the M.A. degree in statistics and the Ph.D. degree in electrical engineering and computer science, both from the University of California, Berkeley, Berkeley, CA, USA, in 2005.

She is currently an Associate Professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. Her current research interests include adaptive and low-power embedded systems design, hardware security, and design intellectual property protection.

Prof. Koushanfar was the recipient of several awards and honors, including the Presidential Early Career Award for Scientists and Engineers, the Association for Computing Machinery Special Interest group on Design Automation (SIGDA) Outstanding New Faculty Award, the National Academy of Sciences Kavli Foundation fellowship, MIT Technology review (MITTR)-35, and the young faculty (or CAREER) awards from the U.S. Army Research Office, the U.S. Office of Naval Research, the Defense Advanced Research Projects Agency, and the National Science Foundation.

Sherief Reda (S'01-M'06-SM'14) received the B.Sc. and M.Sc. degrees from Ain Shams University, Cairo, Egypt, in 1998 and 2000, respectively, and the Ph.D. degree in computer science and engineering from the University of California, San Diego, San Diego, CA, USA, in 2006.

He is currently an Associate Professor with the School of Engineering, Brown University, Providence, RI, USA. His current research interests include energy-efficient adaptive computing for computing systems, thermal/power sensing and modeling for integrated circuits, and low-power VLSI CAD techniques.

Prof. Reda was the recipient of the Best Paper Award from the DATE'02 and the International Symposium on Low Power Electronics and Design (ISLPED)'10, and the National Science Foundation CAREER Award in 2010.