<table>
<thead>
<tr>
<th>Title</th>
<th>Detecting hardware trojan through time domain constrained estimator based unified subspace technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Xue, Mingfu; Liu, Wei; Hu, Aiqun; Wang, Youdong</td>
</tr>
<tr>
<td>Date</td>
<td>2014</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/19687">http://hdl.handle.net/10220/19687</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2014 The Institute of Electronics, Information and Communication Engineers. This paper was published in IEICE Transactions on Information and Systems and is made available as an electronic reprint (preprint) with permission of The Institute of Electronics, Information and Communication Engineers. The paper can be found at the following official DOI: <a href="http://dx.doi.org/10.1587/transinf.E97.D.606">http://dx.doi.org/10.1587/transinf.E97.D.606</a>. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper is prohibited and is subject to penalties under law.</td>
</tr>
</tbody>
</table>
Detecting Hardware Trojan through Time Domain Constrained Estimator Based Unified Subspace Technique*

Mingfu XUE†,‡‡, Student Member, Wei LIU†, Aiqun HU†, and Youdong WANG†, Nonmembers

SUMMARY Hardware Trojan (HT) has emerged as an impending security threat to hardware systems. However, conventional functional tests fail to detect HT since Trojans are triggered by rare events. Most of the existing side-channel based HT detection techniques just simply compare and analyze circuit’s parameters and offer no signal calibration or error correction properties, so they suffer from the challenge and interference of large process variations (PV) and noises in modern nanotechnology which can completely mask Trojan’s contribution to the circuit. This paper presents a novel HT detection method based on subspace technique which can detect tiny HT characteristics under large PV and noises. First, we formulate the HT detection problem as a weak signal detection problem, and then we model it as a feature extraction model. After that, we propose a novel subspace HT detection technique based on time domain constrained estimator. It is proved that we can distinguish the weak HT from variations and noises through particular subspace projections and reconstructed clean signal analysis. The reconstructed clean signal of the proposed algorithm can also be used for accurate parameter estimation of circuits, e.g., power estimation. The proposed technique is a general method for related HT detection schemes to eliminate noises and PV. Both simulations on benchmarks and hardware implementation validations on FPGA boards show the effectiveness and high sensitivity of the new HT detection technique.

key words: information security, hardware security, hardware Trojan detection, unified subspace technique, time domain constrained estimator

1. Introduction

In recent years, the malicious alteration of integrated circuits (IC), also referred to as hardware Trojan (HT), has become an emerging security threat in the hardware community [1]. HT can make the IC malfunction, leak confidential information, or lead to other catastrophic consequences, and thus it has raised serious concerns from industry, military and other critical communities. HT detection techniques are urgently needed to ensure trust in ICs. However, HT detection is extremely difficult [1].

Authors in [2] proposed an approach using multiple excitations of rare logic conditions to increase the HT detection probability of logic testing. This logic testing approach has difficulty in triggering large Trojans that have complicated triggering conditions. The side-channel signal analysis techniques [3], [4] are effective in extracting Trojan signals by monitoring delay, leakage power, and supply current of the circuit. However, these methods cannot deal with the process variation (PV) and the environment noise, which are significantly increasing in ICs fabricated using modern nanotechnology. Moreover, these methods which analyze global signals cannot scale well to large circuits. Therefore, a few localized signal analysis approaches have been proposed to magnify Trojan’s contributions to the circuit [5], [6].

Most of the existing side-channel based HT detection techniques just simply analyze circuit’s parameters and offer no signal calibration or error correction properties. Thus they suffer from the interference of large PV and noises. PV and noises can completely mask Trojan’s contribution to the circuit, and thus allow HT to escape detection.

This paper presents a novel HT detection method based on subspace technique which aims at detecting tiny HT characteristics under large PV and background noise. First, we formulate the HT detection problem as a feature extraction model. Then, we propose a time domain constrained estimator (TDCE) based subspace technique for HT detection.

Subspace techniques have already been used in other applications. In this work, we develop a new subspace technique, including new criterion, objective function, constraints and processes, for HT detection. To the best of our knowledge, this is the first unified subspace HT detection technique. We note that the authors in [3] mentioned a few concepts of subspace and projections for HT detection. However, what they really use is traditional Karhunen-Loeve (KL) expansion and eigenvalue spectrums in the entire space, while our approaches are focusing on individual component, e.g., $x_{d, p}$ principal components. Moreover, the major drawback of [3] is that their traditional subspace method can’t overcome PV and noises. When the HT are small, their subspace method is ineffective or failed. Our subspace technique are essentially different to [3]. In our work, the major novelty is exploring a new subspace technique for the specific purpose of HT detection, which can detect small Trojans under large PV and noises. Our approaches including de-noising, individual component analysis in subspace domain, and reconstructing clean signal.

Both the simulation experiments in ISCAS89 benchmark circuits and hardware implementation validation on FPGA boards show that by applying the proposed scheme, the tiny Trojans which escape the original side-channel detection methods can now be detected. Indeed, the proposed method can provide a general technique for related HT de-
tection methods to eliminate the huge negative impacts of PV and noises. Moreover, the reconstructed clean signal can also be applied for accurate parameter estimation of circuits.

2. Proposed Subspace HT Detection Technique

We consider the HT detection problem as a weak signal detection problem and solve it using the following feature extraction model: For the measured noisy signal, extracting the desired features for Trojan detection, while suppressing all other interferences caused by PV or noises. This procedure is equal to a filtering operation on the measured noisy signal, while the filter is the subspace operator. After this filtering operation, we can obtain the clean signal estimation.

The overall flow of the proposed subspace HT detection scheme is as follows. The input is the target signal from continuous measurements of a circuit under authentication (CUA). The signal will be further sampled and divided into different time windows. Then the TDCE is proposed to divide the target signal into subspaces under certain de-noising parameter constraints and objective functions. After that, we use subspace projection and principal component projection to analyze the subspaces data for HT detection. TDCE is also used to reconstruct the clean signal which can further be used for HT detection and parameter estimation.

2.1 Time domain constrained estimator

The measurements of the CUA can be given by:

\[ y = x + w = \sum_{m=1}^{M} s_m v_m + w = Vs + w \]  

(1)

where \( y \in \mathbb{R}^{K \times 1} \) is a column vector of the measured noisy signal, \( x \in \mathbb{R}^{K \times 1} \) is the clean signal we want to estimate, \( w \in \mathbb{R}^{K \times 1} \) is a K-dimensional vector of noise and variations, and \( V = \{v_1, v_2, \ldots, v_M\} \) is a \( K \times M \) matrix with \( v_m \in \mathbb{R}^{K \times 1} \). Then the covariance matrix of \( y \) is

\[ R_y = E[yy^T] = R_x + R_w \]  

(2)

Let \( \hat{x} = Hy \) be a linear estimator of \( y \) where \( H \) is a \( K \times K \) matrix. The residual signal obtained in this estimation is

\[ r = \hat{x} - x = Hy - x = H(x + w) - x \]

\[ = (H - I)x + Hw \triangleq r_s + r_w \]  

(3)

where \( r_s \triangleq (H - I)x \) represents signal distortion and \( r_w = Hw \) represents the residual noise. Let

\[ \varepsilon_s^2 \triangleq \text{tr}\left( E[r_s r_s^T] \right) = \text{tr}\left( (H - I)x (x^H)(H - I)^H \right) \]

\[ = \text{tr}\left( (H - I)R_s(H - I)^H \right) \]  

(4)

be the energy of the signal distortion vector \( r_s \). Let

\[ \varepsilon_w^2 \triangleq \text{tr}\left( E[r_w r_w^T] \right) = \text{tr}\left( HE \left[ ww^T \right] H^T \right) \]

\[ = \text{tr}\left( HR_w H^T \right) = \sigma_w^2 \text{tr}\left( HH^T \right) \]  

(5)

denotes the energy of the residual noise vector \( r_w \). The objective function and constraints of the TDCE are:

\[ \min_{H} \varepsilon_s^2 \quad \text{Subject to:} \quad \frac{1}{K} \varepsilon_w^2 \leq \alpha \sigma_w^2 \]  

(6)

where \( 0 \leq \alpha \leq 1 \). The estimator minimizes the signal distortion over all linear filters with permissible residual noise level \( \alpha \sigma_w^2 \).

In what follows, we describe how to solve the above TDCE equation. The optimal estimator of (6) can be found using the Kuhn-Tucker necessary conditions for the constrained minimization.

\[ L(H, \mu) = \varepsilon_s^2 + \mu \left( \varepsilon_w^2 - \alpha \sigma_w^2 \right) \]  

(7)

and \( \mu \left( \varepsilon_w^2 - \alpha \sigma_w^2 \right) = 0 \), for \( \mu \geq 0 \).

From \( \frac{\partial L(H, \mu)}{\partial H} = 0 \) we obtain

\[ H_{tdce} = R_s (R_s + \mu \sigma_w^2 I)^{-1} \]  

(9)

The optimal filter (9) is a Wiener filter with an adjustable input noise level \( \alpha \sigma_w^2 \). Applying the eigenvalue decomposition of \( R_s \) to (9), we can rewrite the estimator as

\[ H_{tdce} = V \left[ \begin{array}{cc} G_{\mu} & 0 \\ 0 & 0 \end{array} \right] V^H = VMG_{\mu}V_M^H \]

\[ = \sum_{m=1}^{M} g_{\mu}(m) v_m v_m^H \]  

(10)

where \( G_{\mu} = \Lambda_M (\Lambda_M + \mu \sigma_v^2 I)^{-1}, g_{\mu}(m) = \frac{\lambda_m}{\lambda_m + \mu \sigma_v^2} \). Therefore, the estimate of the clean signal \( x \) is \( \hat{x} = H_{tdce} y \).

2.2 Subspace projections and related terms

From the above TDCE procedure, we have already got:

\[ R_X = [U_s, U_n] \left[ \begin{array}{cc} \Sigma_s & 0 \\ 0 & \Sigma_n \end{array} \right] \left[ \begin{array}{c} U_s^H \\ U_n^H \end{array} \right] = U_s \Sigma_s U_s^H + U_n \Sigma_n U_n^H \]  

(11)

In which, \( U_s \triangleq [u_1, u_2, \ldots, u_r] \), \( U_n \triangleq [u_{r+1}, u_{r+2}, \ldots, u_n] \), \( \Sigma_s = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r) \), \( \Sigma_n = \text{diag}(\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_n) \). \( \lambda_1, \lambda_2, \ldots, \lambda_r \) are called principal eigenvalues, while \( \lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_n \) are called minor eigenvalues.

Definition 1: \( U_s \) is defined as the signal subspace corresponding to the principal eigenvalues. \( U_n \) is defined as the noise subspace corresponding to the \( n-r \) minor eigenvalues. The projection matrix of \( U_s \) is defined as \( P_s \triangleq U_s U_s^H \), and the projection matrix of \( U_n \) is defined as \( P_n \triangleq U_n U_n^H \). Given any matrix \( B \), the projection of \( B \) to \( P_s \) is \( U_s U_s^H B \), and the projection of \( B \) to \( P_n \) is \( U_n U_n^H B \).

3. Simulation Evaluation

We design and insert a sequential Trojan in ISCAS89 benchmark for evaluation. The Trojan is a 3 bit counter, as shown
in Fig. 1. The output of the HT is left unconnected to avoid affecting the circuit’s logic outputs (HT needs to be stealthy which can’t affect the outputs of the circuit for most inputs). It is not a real HT since it doesn’t have any practical HT functionality. The Trojan is extremely tiny occupying 0.006% of the total circuit area of s38417. The PV is set to be 5%. The designs are synthesized with 65 nm TSMC technology. Synopsys PrimePower is used for power analysis and HSPICE is used for extracting circuit level parameters.

It’s observed that we cannot discover this HT from the time domain power trace due to the PV’s effect is far greater than the Trojan’s contribution to the circuit’s parameters. Then, we apply the proposed technique. Figure 2 gives the projections of power trace to the first principal component of the signal subspace. The projection is calculated according to Definition 1. Figure 2 shows that Trojan-inserted CUA has a significant difference from the golden chip while Trojan-free CUA matches well with the golden chip, which indicates the existence of the Trojan. Similar results can be obtained when power trace is projected to other principal components (2nd, 3rd, etc.). However, we found that the most obvious difference between Trojan-inserted CUA and golden CUA achieved when we projected power trace to the 1st principal component, in which, the curves’ shapes are like arch with no crossover points. When projected to 2 or more principal components, the curves become irregular/wavy (See Fig. 6 in the hardware implementation experiments). Indeed, the 1st principal component mainly reflects the characteristics of the circuit itself, while using more principal components will introduce more impurities and interference.

Table 1 gives the comparison of average percentage difference of the time domain signal analysis and the proposed subspace HT detection technique. Column 2 is the average percentage difference achieved by the time domain signal analysis. Since the PV is set to be 5%, these tiny data anomalies between the CUA and the golden chip are completely masked. Therefore the HT cannot be detected in time domain. Column 3 is the average percentage difference achieved by the proposed technique. Although some data anomalies are still smaller than the PV, the proposed technique is proved to have a strong detection capability and high detection sensitivity taking into account of the size of the tiny HT. Real HT which has practical malicious functions must have a much bigger proportion than this sample HT, thus can be detected using the proposed technique. In fact, the peak percentage difference of the proposed technique from some specific time windows is much higher than the PV level, so all of the HT can be detected using the proposed technique.

Figure 3 gives the reconstructed clean power signal analysis for HT detection. Obviously, the reconstructed power signal differential analysis can also reveals the HT.

4. Hardware Implementation Validation

To evaluate our HT detection technique under real PV and noises, hardware validation was also performed using FPGA platforms. The FPGA was Altera CYCLONE IV E EP4CE15F17C8. To obtain the value of power consumption, we measured the voltage drop across a sense resistor (1 Ω), while a differential probe is used to measure the voltage waveforms. These current and voltage waveforms are recorded using a LeCroy waveRunner 6100A oscilloscope.

The design mapped into FPGA chips is a hardware
encryption system using Advanced Encryption Standard (AES) encryption for the software Tencent QQ [7]. We design a HT to the system leaking confidential information. The experimental design and the HT logic are shown in Fig. 4 and Fig. 5. The HT writes the plaintext and the key of the AES to a specific memory, and dumps the data to the bus. Then, malware software or a probe can access confidential information stealthily. The trigger of the HT is the input of the AES module. The HT occupies 0.873% of the total gates of the design. However, since we monitor the power of the entire test board, not only the FPGA, the HT occupies much smaller proportion of the system than this ratio.

Both the golden design and the Trojan-inserted design are mapped into the two FPGA boards in succession. The same input (a long period of voice data) is applied to the two boards. Then the total power consumptions of the two boards are measured. First, we analyze the time domain power data. For the same test board, the power consumption percentage difference of the Trojan-inserted CUA and the golden chip is only 0.032762%, which is negligible and will be certainly masked by the PV of different boards and the measurement noise. Thus, the time-domain power analysis can’t detect this HT. Afterwards, we apply the proposed technique. As shown in Fig. 6, the power traces are projected to the two principal components of the signal subspace. It is obvious that even for this tiny HT under real large PV and noises, the proposed technique can detect the HT.

Note that, the curves in Fig. 2 are smoother than that in Fig. 6, which might be caused by more noises in real measurement than that in simulation experiments, and the time window in Fig. 6 is much longer than that in Fig. 2 thus introduces more fluctuations. For different time windows, the location of Trojan-inserted CUA relative to the Trojan-free CUA changes. This means the Trojan-inserted CUA may locate below (Fig. 6) or above (Fig. 2) the Trojan-free CUA. The Trojan-inserted CUA curve may even has crossover points with the golden chip curve. Whatever, the Trojan-inserted CUA always has significant difference with golden chips and the Trojan-free CUA.

5. Conclusion

We presented a novel subspace HT detection technique which can detect tiny HT under large PV and noises. The main advantage of this technique compared to the state-of-the-art alternatives is that this method can address the PV and noises. The reconstructed clean signal of the proposed technique can also be used for clean parameter estimation.

References