FTAI: Frequency-based Trojan-Activity Identification Method for 3D Integrated Circuits

(Invited Paper)

Zhiming Zhang and Qiaoyan Yu
Department of Electrical and Computer Engineering, University of New Hampshire
Durham, NH 03824, USA

***Abstract*—[1] Three-dimensional (3D) integration emerges as a promising solution to enable technology further scaling. However, 3D chip fabrication and assembly involve more steps than the manufacturing process for conventional 2D integrated circuits (ICs). Recent literature indicates that the extended supply chain for 3D ICs could bring in new security threats, such as 3D hardware Trojans. This work proposes a Frequency-based Trojan-Activity Identification (FTAI) method to detect 3D hardware Trojans. The FTAI method is capable of differentiating the frequency changes induced by Trojans from those caused by process variation noise. Our case study indicates that the proposed peak distance metric is over 30× higher than the Euclidean distance used in the existing literature. Theoretical analysis and simulation results show that the proposed method can tolerate more noise than the time-domain Trojan detection method and thus improve the Trojan detection rate by 38.1%.**

***Index Terms*—Three-dimensional integrated circuit (3D IC), hardware Trojan, Trojan detection, process variation.**

I. Introduction

As technology scaling is approaching its physical limits [1], three-dimensional (3D) integration has emerged as an alternative way to further advance chip density by stacking multiple 2D dies vertically. However, 3D integration techniques could bring in new security threats, such as 3D hardware Trojans [2]. Aiming at altering the original logic or leaking information, hardware Trojan attacks can be performed in any phase of the IC design and fabrication phases. As the fabrication of 3D ICs may involve multiple foundries for die and through-silicon-via (TSV) manufacturing and die-to-die bonding, the extended supply chain could create more Trojan attack surfaces.

Existing Trojan-detection methods are mostly proposed for conventional 2D ICs. Their detection effect might be degraded in a 3D environment. For example, functional-verification based methods may not work well in 3D scenarios. First, the larger number of transistors integrated into the 3D package makes the exhaustive functional verification more sophisticated and time-consuming. Moreover, the limited probing capabilities do not allow us to simultaneously access all die-to-die vertical communication channels for thorough testing neither. Side-channel based Trojan detection is commonly used in securing 2D ICs. However, because 3D ICs usually have more internal noise than 2D ICs, the signal-to-noise ratio (SNR) of the side-channel signals for detection will be reduced noticeably. Larger variations on the process, voltage, and temperature in 3D ICs further lead to a higher false-positive rate. Thus, it is more challenging to precisely extract Trojan's impact on side-channel indicators or/and functional behaviors in the 3D scenarios [3].

To facilitate side-channel based Trojan detection in 3D ICs, it is imperative to develop a new method to tolerate the interference from 3D noise. In this work, we propose a Frequency-based Trojan-Activity Identification (FTAI) to detect 3D Trojans. Our FTAI method tolerates 3D noise and achieves a high Trojan detection rate. Comparing to the existing frequency-based detection methods, such as [4], FTAI takes process variation into consideration and provides a new way of threshold generation without using a fabricated golden chip. Our theoretical analyses verify that the Trojan effect is more differentiable in the frequency spectrum than in timing waveform, no matter it acts as an additive or multiplicative noise. The experimental results further show that FTAI increases the Trojan detection rate by 38.1% compared to the time-domain detection method.

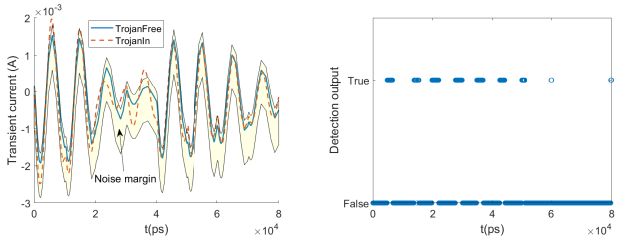
II. Trojan Model

3D hardware Trojans are characterized in the recent work [2]. In this work, we aim to detect the cross-tier hardware Trojan in 3D ICs. The goal of the 3D Trojan is to leak the secret key of a crypto unit implemented in the middle tier. The trigger is located in the same tier as the crypto unit while the payload is in the top tier. The trigger and payload circuits are inserted in two different single-die fabrication phases. According to the cross-tier hardware Trojan model in [2], the Trojan is not functioning during the single-tier testing stage but will be triggered after all 3D tiers are assembled.

We extend the MOLES Trojan [5], which is modeled for 2D ICs, to a 3D version. MOLES is composed of a set of registers as a ring generator to produce a series of random numbers, which will be XORed with the crypto key. The XOR outputs drive a set of capacitors as the Trojan payload. Attackers who know the implementation details of the ring generator can decode the obfuscated key information via power analysis. However, the power consumed in the load capacitors seems like noise if the random sequence is unknown. To form a cross-tier Trojan, the trigger and the ring generator of MOLES are inserted in the middle tier of our transistor-level 3D chip. The crypto unit, an AES Sbox, is located in the middle tier as well. The crypto key for AES will be leaked with eight capacitors. More details are available in Section V-A.

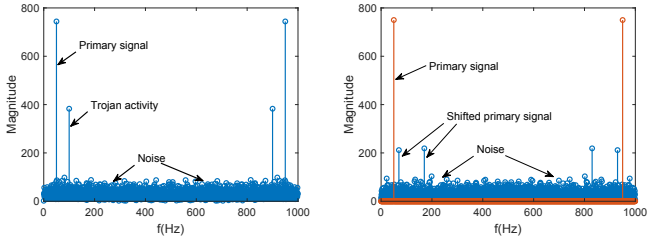
[1]This work is partially supported by the NSF grants No.1652474 and No.1717130.

978-1-7281-8058-8/20/$31.00 ©2020 IEEE



(a) (b)

Fig. 1. Time-domain analysis for Trojan detection. (a) Transient currents for three test cases, (b) Success/Failure of Trojan detection.



(a) (b)

Fig. 2. Frequency spectrum of (a)I_{tot} and (b) $I_{tot_{mul}}$.

III. Limitation of Time-domain Analysis based Trojan Detection

Time-domain analysis on the transient current of the circuit under Trojan attacks could reveal the presence of hardware Trojans, which contribute to more/less current. The efficiency of time-domain analysis heavily depends on the difference between the Trojan-induced current change and pre-existing inherent noise. A smaller difference leads to a higher false-positive/negative detection rate. Figure 1(a) shows the timing waveform for the transient current measured from our transistor-level 3D chip. The current was collected from the power-supply pin of the chip. The *TrojanFree* line in the graph represents a basic 3D chip. The *TrojanIn* line indicates the current after the injected Trojan is triggered. We further introduced process variation to the TrojanFree case to create noise margins, which are highlighted by the yellow area. As shown in Fig. 1(a), the impact of Trojans on the transient current does not exceed the boundaries defined by the noise for most of the time. If we consider the cases in which the TrojanIn line goes beyond the noise margin as the success of Trojan detection, the detection output is shown in Fig. 1(b). A very small portion of the line reaches *True* (i.e., detected) and the overall success detection rate is only 16.98%.

IV. Proposed Frequency-based Trojan-activity Identification (FTAI) for 3D Hardware Trojans

A. Overview of Proposed FTAI Method

As 3D integration techniques bring in new security threats to ICs, it is imperative to develop effective Trojan detection methods for 3D chips. Since time-domain Trojan analysis methods suffer from noise interference, we explore new methods performed in the frequency domain. In this work, we propose a frequency-based Trojan-activity identification (FTAI) method, which exploits the frequency spectrum of the transient current of a 3D chip under Trojan attacks to detect hardware Trojans. We follow the footprint of the work [4] but specifically tailor the detection method for 3D ICs, which are known to have more variation on process/voltage/temperature and internal noise. Different than the work [4], our method waives the assumption on the frequency band of potential Trojans and the independence between primary circuits and Trojans. Furthermore, we propose a new threshold generation algorithm to achieve a high Trojan detection rate and reduce the false-positive rate over the existing work.

B. Theoretical Analysis

First, we assume that I_{Prime} and I_{HT} represent for the transient current contributed by the primary circuit and hardware Trojan, respectively. If the Trojan is an extra circuit that is independent of the primary circuit (i.e., victim module), we can model the total current for the circuit suffering from the Trojan attack with the expression shown in Eq. (1).

$$\begin{aligned} I_{tot} &= I_{Prime} + I_{HT} + n(t) \\ &= A_{Prime} sin(2\pi f_{Prime} t) + A_{HT} sin(2\pi f_{HT} t) + n(t) \end{aligned} \tag{1}$$

In which, A_{Prime} and f_{Prime} represent the amplitude and frequency for I_{Prime}. Similarly, A_{HT} and f_{HT} are the amplitude and frequency of the Trojan current I_{HT}. We use sinusoidal functions to model the current components since most kinds of signals in nature can be modeled with a format of sinusoids [6]. The term $n(t)$ is white noise.

After *Fourier Transformation*, we will observe that the frequency spectrum $\mathcal{F}(I_{tot})$ includes three kinds of frequency components as shown in Eq. (2). Because the frequency response of the white noise is a constant value approximately, we use C to substitute $\mathcal{F}(n(t))$. The corresponding spectrum for $\mathcal{F}(I_{tot})$ is shown in Fig. 2(a). Different than the frequency response of noise, which is flat at the bottom of the entire spectrum, the Trojan activity will result in unique and substantial frequency response.

$$\begin{aligned} &\mathcal{F}(I_{tot}) \approx \frac{A_{Prime}}{2i}[\delta(f - f_{Prime}) + \delta(f + f_{Prime})] + \\ &\frac{A_{HT}}{2i}[\delta(f - f_{HT}) + \delta(f + f_{HT})] + C \end{aligned} \tag{2}$$

In addition to the additive influence on the total current, the current contribution from the hardware Trojan can be modeled as a multiplicative component if the Trojan is inserted by performing malicious modifications to the primary circuit. We formulate the total current $I_{tot_{mul}}$ in Eq. (3). After performing the Fourier transformation on $I_{tot_{mul}}$, we can obtain the frequency-domain expression for the total current $\mathcal{F}(I_{tot_{mul}})$, which is expressed in Eq. (4).

$$I_{tot_{mul}} = (I_{Prime} \times I_{HT}) + n(t) \tag{3}$$

$$\begin{aligned} &\mathcal{F}(I_{tot_{mul}}) \approx \frac{A_{Prime} A_{HT}}{-4}\{\delta[f - (f_{Prime} + f_{HT})] + \\ &\delta[f - (f_{Prime} - f_{HT})] + \delta[f - (-f_{Prime} + f_{HT})] + \\ &\delta[f - (-f_{Prime} - f_{HT})]\} + C \end{aligned} \tag{4}$$

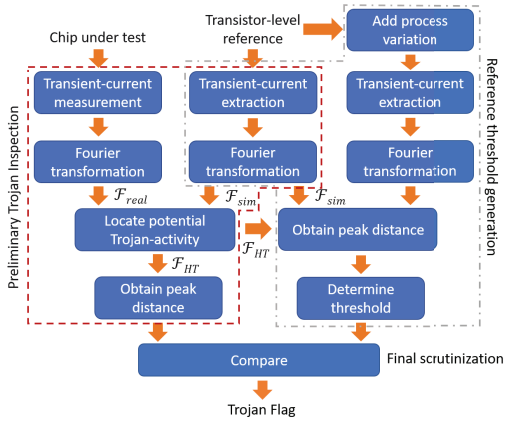


Fig. 3. Trojan detection flow proposed in our FTAI method.

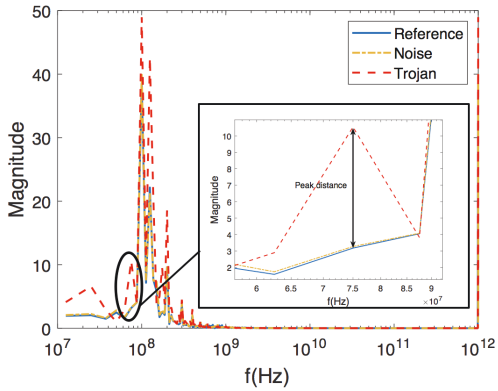


Fig. 4. Comparison of frequency spectra for baseline, noisy, and Trojan impacted cases.

Because multiplication in the time domain is transformed to convolution in the frequency domain, the frequency of the primary current will have an offset induced by the Trojan. Figure 2(b) shows the spectra of $\mathcal{F}(I_{tot_{mul}})$ and primary signal together. We can see the frequency of the primary signal is shifted by the Trojan. In conclusion, our theoretical analysis indicates that the impact of Trojans on the frequency spectrum can be easily differentiated from white noise. This motivates us to propose a frequency-based detection method for 3D Trojans.

C. Detection Flow

The detection flow for the proposed FTAI method is composed of three phases: *preliminary Trojan inspection*, *reference threshold generation*, and *final scrutinization*. Figure 3 depicts the detailed detection flow.

In the phase of the preliminary Trojan inspection, one needs to collect the total transient current of the 3D chip from the power-supply pin. Then, *Fourier transformation* is utilized to convert the time-domain current trace to its frequency-domain representation $\mathcal{F}_{real}$. Next, the same process is repeated on the transistor-level 3D model for the same 3D chip to obtain $\mathcal{F}_{sim}$. The two frequency spectra $\mathcal{F}_{real}$ and $\mathcal{F}_{sim}$ are compared to identify the suspicious frequency band $\mathcal{F}_{HT}$, in which the Trojan may be located. This process will minimize the noise interference on Trojan detection, as discussed in Section IV-B. We performed a simulation to compare the frequency spectra for the current trace of the golden model (i.e., clean without noise and Trojan), noisy model (i.e., noise induced by process variation is considered), and Trojan-infected model (i.e., the triggered Trojan leaks information). As shown in Fig. 4, the Trojan results in a new frequency peak on the lower frequency band than the primary signal. The zoom-in view of that frequency peak indicates that the Trojan introduces a more substantial magnitude difference than the process variation noise. To facilitate Trojan scrutinization in the following phases, we define a metric, named *peak distance (PD)*, to quantify the difference in frequency magnitude between the first frequency peak induced by the Trojan and the corresponding response from the reference model.

After the preliminary Trojan inspection, a reference threshold will be applied to further examine the suspicious frequency band. As golden chips are often unavailable in practical situations, the reference threshold is provided based on simulations [7]. In this work, we apply different process variations to our transistor-level reference and obtain the corresponding frequency spectra. In each spectrum, we measure the peak distance against $\mathcal{F}_{sim}$ in the frequency band $\mathcal{F}_{HT}$. We denote the group of the peak-distance values for all the cases as PD_{noise}. It is used to evaluate the magnitude changes induced by noise on $\mathcal{F}_{HT}$. To achieve a high confidence, we apply the 3σ value of the signal PD_{noise} as the threshold PD_{th} to our Trojan detection method. The closed-form expression for PD_{th} is available in Eq. (5), where μ and σ are the mean and the standard deviation of PD_{noise}.

$$PD_{th} = \mu + 3\sigma = Mean(PD_{noise}) + 3Std(PD_{noise}) \quad (5)$$

In the phase of final scrutinization, the *peak distance* of the chip under examination is compared with the threshold generated in the previous phase. If the peak distance exceeds the given threshold, we conclude that there is a Trojan inserted in the chip.

V. Experimental Results

A. Experimental Setup

We evaluated the proposed method by using transistor-level simulations. A stacked 3D IC with three tiers is implemented in a 45nm NCSU FreePDK technology [8]. The PDN in each tier is mainly composed of a global power grid and a virtual grid. The local load circuits in each tier are multiple inverters. In our experiments, the target of the MOLES Trojan (described in Section II) is an AES Sbox implemented at transistor level. We provided the input vectors satisfying the triggering condition of 3D MOLES Trojans to leak the crypto key during our experiments. We collected the transient current trace for a period of 80ns from the power-supply pin of the transistor-level 3D chip and converted the time-domain current traces to frequency spectra. We repeated the same procedure for the models of Trojan-free (i.e., *reference*), Trojan-free but considering different process variation noise (i.e., *noise*), Trojan-injected at the nominal process variation (i.e. *Trojan*), and Trojan-injected in different process variation cases.

TABLE I
TROJAN DETECTION METRICS USED IN FREQUENCY-DOMAIN AND TIME-DOMAIN ANALYSIS METHODS.

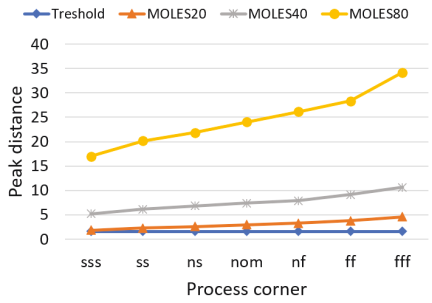
Metrics w.r.t Trojan size	MOLES20	MOLES40	MOLES80
Euclidean distance	0.0899	0.1831	0.6049
Proposed peak distance	2.986	7.367	24.007
Improved distance	33.2×	40.2×	39.7×

B. Impact of Trojan size on Trojan Detection

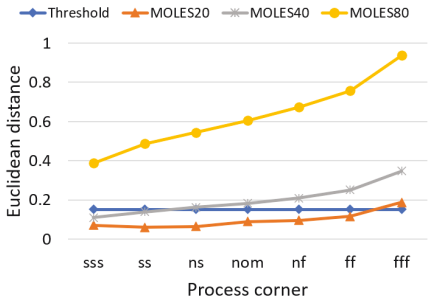
All the experiments in this subsection were based on the nominal process variation. We first compared the proposed detection metric *peak distance* in the frequency domain, with *Euclidean distance* in the time domain. We performed the proposed spectrum analysis and identified the Trojan-related frequency peak in 75MHz. Peak distance for three Trojan sizes (MOLES20, MOLES40, and MOLES80) was measured. MOLES20 means that there are 20 registers in the MOLES ring generator. As shown in Table I, the proposed peak distance is always 30× higher than Euclidean distance. This means the proposed frequency-domain analysis method can better tolerate the measurement errors and noise interference than the time-domain Trojan detection. We applied the seven process corners to the reference chip and collected their corresponding *peak distance* to form the group PD_{noise}. After following the procedure introduced in Section IV-C, we obtained its 3σ value of 1.578 for our frequency-domain Trojan analysis. As all measured peak distance values are greater than 1.578, our method can detect all three Trojans. In contrast, the 3σ value for the time-domain Trojan analysis is 0.1521, which is higher than the Euclidean distance for MOLES20. Thus, the time-domain Trojan detection fails in the MOLES20 case.

C. Impact of Process Variation on Trojan Detection

We further evaluated the impact of process variation on the Trojan detection success rate of our method. We conducted different test cases by applying seven process variation configurations to our 3D structure. The seven corners are sss (the slowest), ss, ns, nom, nf, ff, and fff (the fastest). The *sss* (*fff*) case doubles the progress variation from *nom* to *ss* (*ff*). The *ns* (*nf*) case is the half variation step from *nom* to *ss* (*ff*). The main variations include the long channel threshold voltage, gate oxide thickness, channel length offset, first-order body effect coefficient, and low-field mobility. As shown in Fig. 5(a), the peak distance of the 3D circuit tampered by Trojans with different sizes is consistently larger than the threshold, which means all the Trojans can be detected and the Trojan detection rate achieved by our method is 100%. The results shown in Fig. 5(b) represent the Euclidean distance obtained by the time-domain analysis method. As can be seen, the Euclidean distance for the case of MOLES20 is always below the threshold (except the fff corner). The time-domain analysis based Trojan detection also fails to detect MOLES40 in the sss and ss cases. We calculated that the time-domain method yields a 61.9% of Trojan detection rate. Thus, our proposed FTAI increases the Trojan detection rate by 38.1%.



(a)



(b)

Fig. 5. Trojan detection effectiveness comparison between (a) frequency-domain method and (b) time-domain method at different noise levels.

VI. CONCLUSION

3D IC is considered as a promising solution for future integration. However, the stacking structure and complicated fabrication process give adversaries a chance to perform malicious attacks. Unexplored 3D Hardware Trojans can be inserted in the supply chain. Very limited works about 3D Trojan's detection and mitigation can be found in the current literature. We proposed an FTAI, which can better tolerate 3D noise than the time-domain detection method to provide a better detection rate on 3D hardware Trojans. The experimental results show that FTAI achieved a 100% detection rate on the 3D-version of MOLES. Comparing to the time-domain method, FTAI improved the detection rate by 38.1%.

REFERENCES

[1] J. Knechtel, O. Sinanoglu, I. A. M. Elfadel, J. Lienig, and C. C. Sze, "Large-Scale 3D Chips: Challenges and Solutions for Design Automation, Testing, and Trustworthy Integration," *T-SLDM*, vol. 10, pp. 45–62, 2017.

[2] Z. Zhang and Q. Yu, "Modeling hardware trojans in 3d ics," in *Proc. ISVLSI'19*, 2019, pp. 483–488.

[3] Y. Xie, C. Bao, C. Serafy, T. Lu, A. Srivastava, and M. Tehranipoor, "Security and Vulnerability Implications of 3D ICs," *TMSCS*, vol. 2, no. 2, pp. 108–122, Apr 2016.

[4] F. Karabacak, U. Y. Ogras, and S. Ozev, "Detection of malicious hardware components in mobile platforms," in *Proc. ISQED'16*, 2016, pp. 179–184.

[5] L. Lin, W. Burleson, and C. Paar, "MOLES: Malicious off-chip leakage enabled by side-channels," in *Proc. ICCAD'09*, Nov 2009, pp. 117–122.

[6] M. M. Goodwin, *Adaptive signal models: Theory, algorithms, and audio applications*. Springer Science & Business Media, 2012, vol. 467.

[7] J. He, Y. Zhao, X. Guo, and Y. Jin, "Hardware trojan detection through chip-free electromagnetic side-channel statistical analysis," *TVLSI*, vol. 25, no. 10, pp. 2939–2948, 2017.

[8] S. M. Satheesh and E. Salman, "Power distribution in tsv-based 3-d processor-memory stacks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 692–703, 2012.