IEEE | IEEE Embedded Systems Letters

Hardware Trojan Detection Method against Balanced Controllability Trigger Design

Journal:	*Embedded Systems Letters*
Manuscript ID	IEEE-ESL-Apr-23-0092.R1
Manuscript Type:	Original Manuscripts
Date Submitted by the Author:	24-Jun-2023
Complete List of Authors:	Hsu, Wei-Ting Lo, Pei-Yu Chen, Chi-Wei Tien, Chin-Wei Kuo, Sy-Yen
Keywords:	embedded systems security, secure hardware
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
gavideo.mp4	

Hardware Trojan Detection Method against Balanced Controllability Trigger Design

Wei-Ting Hsu, Pei-Yu Lo, Chi-Wei Chen, Chin-Wei Tien, Sy-Yen Kuo

***Abstract*—Hardware Trojan (HT) has become a serious threat to the Internet of Things due to the globalization of the integrated circuit industry. To evade functional verification, HTs tend to have at least one trigger signal at the gate-level netlist with a very low transition probability. Based on this nature, previous studies use imbalanced controllability as a feature to detect HTs, assuming that signals with imbalanced controllability are always accompanied by low transition probability. However, this study has found out a way to create a new type of HT that has low transition probability but balanced controllability, against previous methods. Hence, current imbalanced controllability detectors are inadequate in this scenario. To address this limitation, we propose a probability-based detection method that uses unsupervised anomaly analysis to detect HTs. Our proposed method detects not only the proposed HT but also the 580 Trojan benchmarks on Trusthub. Experimental results show that our proposed detector outperforms other detectors, achieving an overall 100% TPR and 0.37% FPR on the 580 benchmarks.**

***Index Terms*—IoT, hardware security, hardware Trojans, controllability, unsupervised clustering, outliers**

I. Introduction

THE semiconductor industry's rapid growth has led to global outsourcing of integrated circuits design and manufacturing. However, outsourcing different design phases can create a threat from Hardware Trojans (HTs), which are a serious security concern for embedded systems.

An HT usually consists of a trigger and a payload, where the activation of the payload can lead to information leakage, functional changes, system failures, and other security risks when the trigger condition(s) are met. This letter focuses on detecting HTs at the gate level using the Third-Party Intellectual Property (3PIP) attack model. It means HTs may be inserted in 3PIPs by malicious vendors and there is no available reference design.

Many techniques have been proposed for detecting HTs at the gate level. One of the well-known methods is COTD [1], which utilizes the Sandia Controllability/Observability Analysis Program (SCOAP) [2] testability analysis and has achieved 100% TPR and 0% FPR on 21 benchmarks published on Trusthub [3], [4]. Since the proposal of COTD, most studies on Trojan detection based on machine learning have adopted testability analysis, especially SCOAP [5]. COTD claimed that Trojan signals with low transition probability should have low testability, which means not only poor controllability but poor observability. Two recent papers [6], [7] have introduced the concept of imbalanced controllability analysis, arguing that lower transition probabilities are always associated with imbalanced 0/1-controllability rather than low testability. Their findings demonstrate that COTD is insufficient for detecting HTs.

In this letter, we pointed out that imbalanced controllability isn't always indicative of lower transition probability. We demonstrated this by constructing HT benchmarks that can evade state-of-the-art imbalanced controllability Trojan detection. Furthermore, we proposed a detection method, which can detect not only our inserted Trojan but also the 580 Trojan benchmarks on Trusthub [3], [4], [8]. Finally, the performance of the proposed detection method is compared among COTD [1] and imbalanced controllability detectors [6], [7] on 580 benchmarks. The experimental results indicated that the proposed detection method does perform better.

II. Background

Previous studies have used various testability analysis methods to detect HTs. This section introduces some of them, which will be compared in this letter.

A. SCOAP Testability Measures

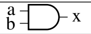
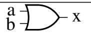
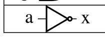
The Sandia Controllability/Observability Analysis Program [2] (SCOAP) is a topology-based testability analysis with linear time complexity. It measures the testability of each net based on several numerical values including combinational 0/1-controllability (CC0/CC1) and combinational observability (CO). The numerical values estimate the minimum signal manipulation required for controlling/observing signals from primary inputs/outputs. The method initializes primary input (CC0, CC1) to (1, 1), computes gate controllability from inputs to outputs using Table I gate propagation rules, and computes observability values from outputs back to inputs using obtained controllability values.

B. COP Testability Measures

The Controllability/Observability Program [9] (COP) is another testability analysis method with linear time complexity. COP has two values C_x and O_x. The C_x is the estimated probability of a signal's value being 1, and the O_x is the estimated probability of fault effect in x being observed at primary outputs. The focus of this paper is on C_x. To calculate the value of C_x, the C_x of primary inputs are first initialized to a value of 0.5. Then, the output value of C_x is determined according to the computation rules specified in Table I. Note that due to the issues of fanout reconvergence, COP analysis doesn't equal actual signal probability, which needs exhaustive simulation.

The authors are with Graduate Institute of Electrical Engineering and Graduate Institute of Communication Engineering, National Taiwan University. E-mail: {r09921a30, r08921a29, r08921a28, cwtien, sykuo}@ntu.edu.tw

TABLE I
GATE PROPAGATION RULES FOR CONTROLLABILITY AND COP

Gate	$CC0(x)$	$CC1(x)$	C_x
	$min\{CC0(a), CC0(b)\}$	$CC1(a) + CC1(b)$	$C_a * C_b$
	$CC0(a) + CC0(b)$	$min\{CC1(a), CC1(b)\}$	$1-(1-C_a)(1-C_b)$
	$CC1(a)$	$CC0(a)$	$1 - C_a$

Note: In the original SCOAP, there should be an additional increment (+1) in each equation listed in Table I. However, it is not considered in the actual values calculated by TetraMAX, which is widely used to get SCOAP values in related studies. For the sake of simplicity and consistency with previous studies, we adopted these calculation rules throughout this letter.

C. Correlation between imbalanced controllability and Transition Probability

The strong correlation between imbalanced controllability and transition probability is found out in [6], [7]. This correlation can be comprehended through the analysis of Fig. 1, where the decreasing difference between P0 and P1 in various circuits corresponds to a decreasing disparity between CC0 and CC1. Given that controllability metrics, such as those obtained from widely-used tools like TetraMax, are easily accessible, imbalanced controllability has been implemented as a feature to identify Trojans. However, this letter highlights imbalanced 0/1-controllability alone is not necessarily indicative of lower transition probabilities due to the distinct meanings of controllability and transition probability. Consequently, Trojans may evade detection by techniques that rely solely on imbalanced 0/1-controllability as a defining characteristic.

III. PROPOSED METHOD

A. HT Insertion Against Imbalanced Controllability

Although imbalanced 0/1-controllability seems to be always accompanied by the imbalanced 0/1-probability, we found out that perfect balanced 0/1-controllability with an imbalanced 0/1-probability is possible. Fig. 2a is an example circuit. If all the circuit's inputs are PIs, the output transition probability would be approximately $1/4$, and the values of Combinational Controllability (CC) are $(CC0, CC1) = (4, 4)$, which are perfectly balanced.

We have shown that a trigger signal with perfectly balanced 0/1-controllability and extremely low transition probability is possible, but it may require too many input signals to maintain the perfect balance. To validate the effectiveness of such triggers in evading detection, we aimed to reduce the degree of imbalance. After inserting a Trojan into four Trusthub benchmarks, we employed two imbalance Trojan detection methods proposed in [6], [7] to detect our inserted benchmarks. The inserted Trojan had 11 signals, which can be primary inputs or any internal signals in the original design. These signals served as inputs for each AND gate in the first level. The other input signals of the rightmost AND gate were constructed similarly but ignored for simplicity, as shown in Fig. 2b. The final trigger probability is approximately $1/2^{32}$, and the degree of imbalanced CC is $44/8 = 5.5$.

The proposed trigger in Fig. 2 is connected to a randomly selected signal to leak secrets through a primary output and is inserted into five circuits. However, concerns arise regarding the validity and observability of such Trojans due to redundant circuitry. To address this, all inserted Trojans are verified using TetraMAX to ensure their ability to propagate the impact to the primary output. Fig. 3 shows the clustering of two imbalanced CC methods [6], [7] on the inserted benchmark "wb_conmax". Both detection methods adopt K-means algorithm (k=3) to classify Trojan signals. The C1 cluster (yellow) is regarded as normal signals and the others (red and blue) are Trojan signals. The only difference between the two methods shows in the features, which are $(CC_0\sqrt{\frac{CC_0}{CC_1}}, CC_1\sqrt{\frac{CC_1}{CC_0}})$ in [6] and $(\frac{CC_0}{CC_1}, \frac{CC_1}{CC_0})$ in [7]. The features of the inserted triggers, which are marked as "X" in Fig. 3, are $(3.4, 103.1)$ in Fig. 3a and $(0.18, 5.5)$ in Fig. 3b. It can be seen that the inserted trigger is classified as a normal signal by both methods. Table II summarizes the detection results for all inserted benchmarks, displaying the centroids of three clusters in columns [3-5] and [8-10]. Notably, all inserted Trojan triggers are classified as the C1 cluster, which represents normal signals. This suggests that the inserted triggers effectively decrease the degree of imbalanced CC while maintaining a low transition probability.

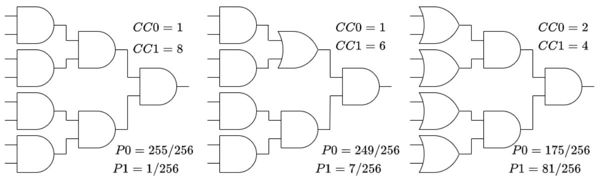


Fig. 1. Examples of the strong correlation between imbalanced controllability and transition probability in [7]. It shows the decreasing difference between P0 and P1 corresponds to a decreasing disparity between CC0 and CC1.

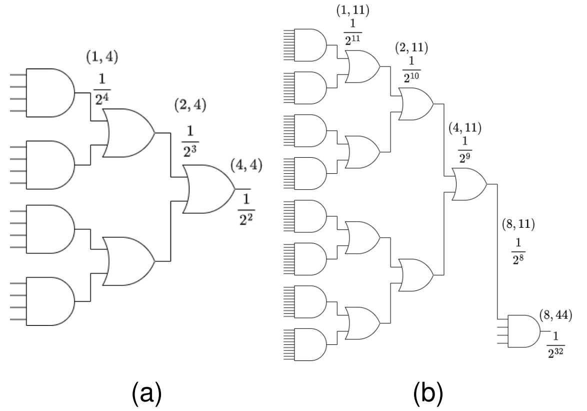


Fig. 2. (a) A circuit with imbalanced transition probability but perfectly balanced controllability. (b) The structure of inserted Trojan. Note that for an OR gate, when the probability of the input signals is the same and represented as $1/2^n$, we simplify the output probability to be $1/2^{n-1}$. It can be easily shown that $P_{approx} - P_{actual} = 1/2^{2n} - 1/2^{2n-1}$, which makes the simplification reasonable when n is not small.

B. HT Trigger Detection Based on COP and CBLOF

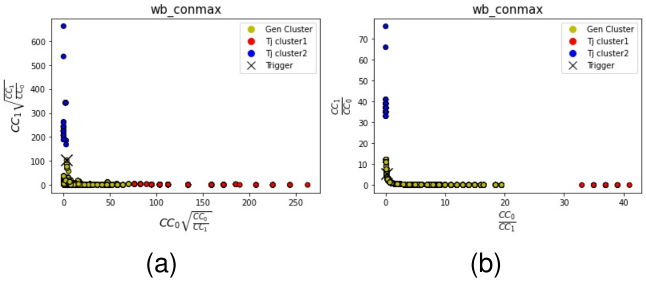
The above discussion shows that the current HT detection methods are insufficient. Therefore, we proposed a reference-free trojan detection method, which utilizes COP testability analysis as a feature and adopts Cluster-Based Local Outlier Factor algorithm [10] (CBLOF) to classify Trojan nets.

COP calculation rules are more correlated to signal probability instead of imbalanced CC. Furthermore, because the

TABLE II
CLASSIFICATION RESULTS OF VARIOUS DETECION METHODS ON INSERTED BENCHMARKS

Bench.	No. of Gate	[6]					[7]					Proposed method	
		C1 (Normal)	C2 (Trojan)	C3 (Trojan)	Trigger's cluster	FPR	C1 (Normal)	C2 (Trojan)	C3 (Trojan)	Trigger's cluster	FPR	Trigger's cluster	FPR
s15850	2155	$(3.7, 4.7)$	$(628.4, 0.2)$	$(0.3, 379.1)$	C1	0.1%	$(1.5, 1.4)$	$(86.0, 0.1)$	$(0.1, 10.8)$	C1	5.2%	Trojan	0.5%
b15	3873	$(18.2, 32.4)$	$(2443.2, 0.3)$	$(0.2, 2484.1)$	C1	0.7%	$(2.9, 3.4)$	$(127.0, 0.1)$	$(0.1, 71.2)$	C1	3.3%	Trojan	0.4%
s38417	5329	$(4.5, 3.9)$	$(292.9, 0.4)$	$(0.9, 232.9)$	C1	1.7%	$(1.5, 1.5)$	$(37.5, 0.1)$	$(0.1, 37.1)$	C1	1.1%	Trojan	0.5%
wb_conmax	20447	$(11.1, 3.1)$	$(134.0, 1.3)$	$(0.8, 261.5)$	C1	1.1%	$(2.7, 1.0)$	$(37.0, 0.0)$	$(0.0, 38.4)$	C1	1.1%	Trojan	0.5%
aes_128	131085	$(11.3, 5.0)$	$(333.6, 2.3)$	$(2.8, 200.8)$	C2	5.5%	$(1.8, 1.2)$	$(8.6, 0.1)$	$(0.2, 7.2)$	C2	22%	Trojan	0%

* Both methods [6], [7] are unable to detect the inserted Trojan in all circuits except for 'aes 128'.
** The circuit types include: 's15850' and 's38417' from the ISCAS'89 benchmarks; 'b15', an 80386 processor from the ITC'99 benchmarks; 'wb_conmax', an interconnect matrix capable of connecting up to 8 masters and 6 slaves; and 'aes_128', an encryption circuit with a 128-bit key.



(a) (b)

Fig. 3. Detection results on inserted Trojan benchmarks wb_conmax using the method proposed in [6], [7]. Both triggers marked as "X" are misclassified as normal signals.

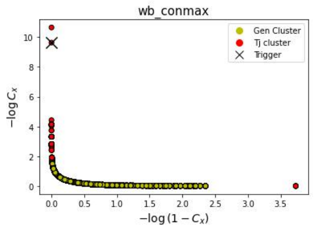


Fig. 4. Detection results on inserted Trojan benchmarks wb_conmax using the proposed method. The trigger marked as "X" is separated far away from the normal signals.

testability analysis rules for COP and SCOAP controllability have great similarity in Table I, we believe the proposed detection method can detect not only Trojan inserted by our insertion method but also the existing Trojan benchmarks on Trusthub. The equation for calculating the COP-based features is

$$(P_0, P_1) = (-\log_2(1 - C_x), -\log_2 C_x) \quad (1)$$

CBLOF, with the time complexity being O(NlogN), is an unsupervised anomaly detection method classifying the data points based on the cluster-based local density of its neighbors. Therefore, the detection method doesn't need golden circuits or labeled data as a reference. In the view of Trojan detection, the extremely low transition trigger should have low C_x, making the data point in (1) of the trigger far away from normal signals. Thus, trigger signals should be considered outliers and classified as Trojan signals.

Fig. 4 depicts the clustering of inserted benchmarks "wb_conmax", where trigger signals are separated further from normal signals than in the imbalanced CC distribution. Using our proposed methodology, all inserted triggers were classified as Trojan signals with a low False Positive Rate (FPR), as shown in column fourteen of Table II. While all methods were able to detect the trigger in "aes_128", our method achieved the lowest FPR. These results indicate that our method is capable of detecting triggers with balanced CC while maintaining a low FPR.

TABLE III
EXPERIMENTAL RESULTS OF VARIOUS DETECTION METHODS

Paper	**Feature**	**Classifier**	**TPR**	**FPR**	**No. of Escaped benchmarks**
[1], Hassan Salmani et al.	$(\sqrt{CC_0^2 + CC_1^2}, CO)$	kMeans	86%	32%	84
[7], Yu Su et al.	$(CC_0\sqrt{\frac{CC_0}{CC_1}}, CC_1\sqrt{\frac{CC_1}{CC_0}})$	kMeans	94%	2%	47
_		CBLOF	98%	0.42%	16
[6], Kai Huang et al.	$(\frac{CC_0}{CC_1}, \frac{CC_1}{CC_0})$	kMeans	97%	7%	21
_		CBLOF	95%	0.46%	17
_	$(-\log(1 - C_x), -\log C_x)$	kMeans	99%	18%	10
Proposed method		CBLOF	100%	0.37%	0

Experimental methods from this letter are denoted as "_"

IV. EXPERIMENTAL RESULTS

The proposed method was applied to 580 TRIT-TC benchmarks [8] to validate it can detect not only Trojans inserted by ourselves in the previous section but the existing Trojan benchmarks on Trusthub. Further, we implement various imbalanced CC detectors [1], [6], [7] to compare the results with the proposed method on the same 580 benchmarks.

A. Experimental setup

To follow the Trojan definition in previous papers [6], [7], we consider only the trigger signal as Trojans. That is, other signals (including the normal signals or signals in the Trojan trigger and payload circuit) classified into Trojan signals are considered false positive signals. The COP features are calculated by the self-developed program. we use the "pyod" package to construct CBLOF classifier and set "contamination" to 0.005. SCOAP values are obtained by TetraMAX.

B. Results and Analysis

Tabel III shows the experimental results of various detection methods [1], [6], [7] and the proposed method. Our method prioritizes security when considering the trade-off between true positive rate (TPR) and false positive rate (FPR). Specifically, we are willing to accept a higher FPR in exchange for achieving a TPR of 100% and minimizing the number of benchmarks that escape from the detector. As a result, designers may need to spend more time manually reviewing the suspicious signals detected by our method. However, it can be seen that the proposed method outperforms the others in terms of TPR and FPR. There are two possible reasons for the result: (1) Imbalanced 0/1-controllability doesn't always accompany by the imbalanced 0/1-probability, and (2) CBLOF has better performance than k-Means.

Reason (1) has been discussed in the previous section. Moreover, Fig. 5b is an example indicating that the proposed COP feature makes the trigger far from normal signals compared to the feature of imbalanced CC in Fig. 5a. Thus, the trigger signal is easier classified as Trojan signals by the COP feature. For reason (2), our findings suggest that the k-Means algorithm is more susceptible to the effects of normal signals with high imbalanced CC in Trojan detection. As shown in Fig. 5a, we observed that normal signals with extremely high imbalanced CC were misclassified as Trojan clusters, leading to the misclassification of trigger signals with slightly lower imbalanced CC as normal signals.

To support our argument, we utilized imbalanced CC and CBLOF on 580 benchmarks. Results in rows [3,5] of Tabel III show improved performance with CBLOF replacing k-Means. However, CBLOF alone did not solely contribute to the positive impact on performance. In row 6, utilizing COP as the feature and K-means clustering as the classifier successfully reduced escaped benchmarks despite a higher FPR. Substituting K-means with CBLOF and using COP as the feature in row 6 further improved the efficacy of our method, outperforming other approaches. Thus, the proposed feature with CBLOF indeed facilitates the classification of trigger signals as abnormal signals more effectively than the imbalanced CC methods.

C. Limitations

The proposed method identifies all trigger signals as Trojans, but it has limitations. While previous research suggests that transition probability is a direct reflection of the stealthy nature of Trojan triggers, this does not mean that observability is irrelevant. In the case of the Trojan type found in the 580 benchmarks and previous papers, which aims to leak secrets by placing its payload near the primary output, observability is low and less important than transition probability. However, for other types of Trojan, observability should be considered along with other factors for effective detection.

D. Performance Analysis in Large-Scale Circuits

Our method's time complexity includes COP with O(N) and CBLOF with O(NlogN), making it ideal for large-scale circuit analysis. We chose to test our method on 'aes 128' in Table II because it is the largest circuit in Trusthub, and our method achieved the lowest FPR. Additionally, our approach may assist test-generation-based Trojan detectors [11], where the time required for functional simulation to get rare nodes may increase with circuit size. Once our method identifies a signal as suspicious, the test-generation-based detectors can strengthen test vector generation for relevant suspicious signals to enhance coverage and reduce runtime in large-scale circuits.

V. Conclusion

This letter examines the correlation between imbalanced controllability and signal probability, highlighting cases where they do not always correlate. Building on this observation, we introduced a new type of Trojan against controllability detectors. We further proposed a detection method, which can detect not only our inserted Trojans but also the existing benchmarks on Trusthub. The proposed approach utilizes an unsupervised anomaly detection technique, which eliminates the need for a golden model. The experimental results on 580 benchmarks show the proposed method outperforms the other imbalanced controllability detectors by achieving an overall 100% TPR and 0.37% FPR.

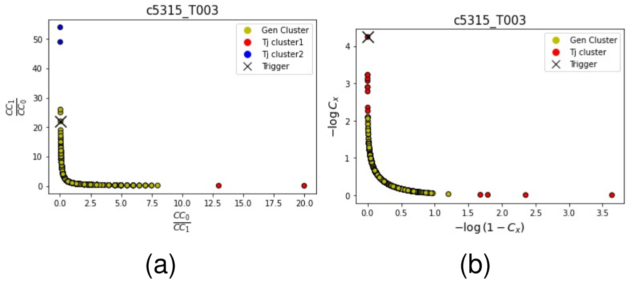


Fig. 5. Detection results on c5315_T003 Trusthub benchmark with different detection methods. The trigger marked as "X" is misclassified as a normal signal in 5a and classified as a Trojan signal in 5b(the proposed method). (a) The results with the method proposed in [7]. (b) The results with the proposed detection method.

References

[1] H. Salmani, "Cotd: Reference-free hardware trojan detection and recovery based on controllability and observability in gate-level netlist," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 2, pp. 338–350, 2016.

[2] L. Goldstein and E. Thigpen, "Scoap: Sandia controllability/observability analysis program," in *17th Design Automation Conference*, 1980, pp. 190–196.

[3] H. Salmani, M. Tehranipoor, and R. Karri, "On design vulnerability analysis and trust benchmarks development," in *2013 IEEE 31st International Conference on Computer Design (ICCD)*, 2013, pp. 471–474.

[4] B. Shakya, T. He, H. Salmani, D. Forte, S. Bhunia, and M. Tehranipoor, "Benchmarking of hardware trojans and maliciously affected circuits," *Journal of Hardware and Systems Security*, vol. 1, no. 1, pp. 85–102, 2017.

[5] P.-Y. Lo, C.-W. Chen, W.-T. Hsu, C.-W. Chen, C.-W. Tien, and S.-Y. Kuo, "Semi-supervised trojan nets classification using anomaly detection based on scoap features," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 2423–2427.

[6] K. Huang and Y. He, "Trigger identification using difference-amplified controllability and dynamic transition probability for hardware trojan detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3387–3400, 2019.

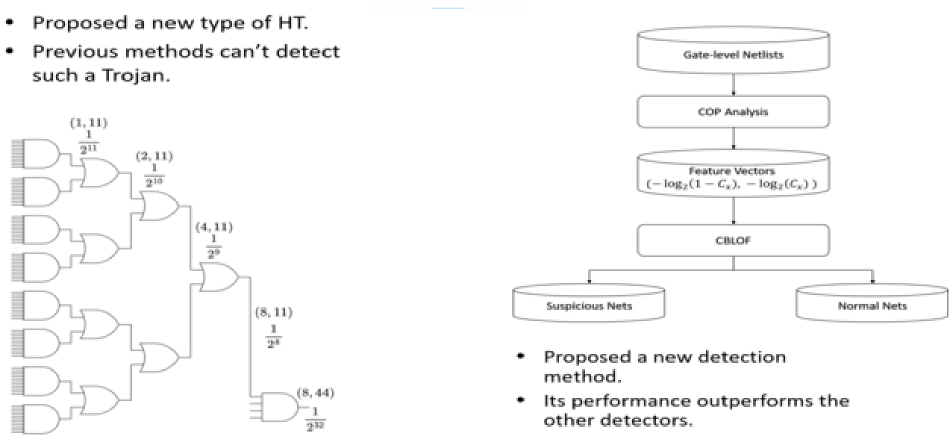
[7] Y. Su, H. Shen, R. Lu, and Y. Ye, "A stealthy hardware trojan design and corresponding detection method," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–6.

[8] J. Cruz, Y. Huang, P. Mishra, and S. Bhunia, "An automated configurable trojan insertion framework for dynamic trust benchmarks," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1598–1603.

[9] F. Brglez, "On testability analysis of combinational circuits," in *Proc. International Symp. Circuits and Systems*, 1984, pp. 221–225.

[10] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern recognition letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[11] R. S. Chakraborty, F. Wolff, S. Paul, C. Papachristou, and S. Bhunia, "Mero: A statistical approach for hardware trojan detection," in *Cryptographic Hardware and Embedded Systems-CHES 2009: 11th International Workshop Lausanne, Switzerland, September 6-9, 2009 Proceedings*. Springer, 2009, pp. 396–410.



- Proposed a new type of HT.
- Previous methods can't detect such a Trojan.

- Proposed a new detection method.
- Its performance outperforms the other detectors.

(1, 11) $\frac{1}{2^{11}}$
(2, 11) $\frac{1}{2^{10}}$
(4, 11) $\frac{1}{2^{9}}$
(8, 11) $\frac{1}{2^{8}}$
(8, 44) $\frac{1}{2^{32}}$
Gate-level Netlists
COP Analysis
Feature Vectors $(-\log_2(1-C_x), -\log_2(C_x))$
CBLOF
Suspicious Nets
Normal Nets

441x197mm (38 x 38 DPI)