

# A Novel Hardware Trojan Insertion Method against SCOAP-based Cluster Detection Method

Chi-Wei Chen, Pei-Yu Lo, Chin-Wei Tien and Sy-Yen Kuo, *Fellow, IEEE*

**Abstract**—The growing specialization in the IC industry has resulted in the outsourcing of IC design and manufacturing to third-party suppliers. Unfortunately, this trend has brought about various hardware security concerns, particularly those related to hardware Trojans. The prevailing gate-level hardware Trojan detection technology still relies on the outliers of Sandia Controllability Observability Analysis Program (SCOAP) value. Previous research focused on generating benchmarks against detectors from scratch, however, in this paper, We are the first to propose a groundbreaking method that focuses on assisting designed and detected hardware Trojans in evading SCOAP-based cluster detection, distinguishing our work from previous studies. Our innovative approach involves strategically inserting simple structures to decrease the SCOAP value of Trojans. This process ensures an upper bound on the trigger probability while expanding the range of possible outcomes. Through rigorous experimentation, we validate the effectiveness of our method in countering SCOAP-based cluster detection. Additionally, we employ the Synopsys TetraMAX tool to demonstrate that our approach does not introduce any new redundancies.

**Index Terms**—Hardware Trojan, Hardware Security, SCOAP

## I. INTRODUCTION

In recent years, there has been an increasing recognition among researchers regarding the significance of hardware security, specifically the threat posed by hardware Trojans (HTs) [5]. HTs involve malicious modifications introduced within integrated circuits (ICs), with attackers aiming to design stealthy HTs and implant them during the IC design and manufacturing process to achieve their objectives [4]. Due to the potential insertion of HTs at various stages of design and manufacturing, the development of detection methods for each stage has become crucial. This letter focuses specifically on gate-level HTs.

The Sandia Controllability Observability Analysis Program (SCOAP) [10] serves as a widely used testability measure in HT detection [9] [7]. It acts as the core method for distinguishing HTs from general circuits, as HTs typically exhibit higher SCOAP scores. However, current research efforts [6] [8] have primarily concentrated on generating HT benchmarks that can evade detection, often with structural constraints limiting the construction of HTs from existing designs. Furthermore, there is a lack of studies addressing the need to assist in countering detection for HTs that have already been detected. Considering the more realistic scenario where HT payloads may have been pre-designed, the development of attack methods specifically targeting SCOAP-based cluster detectors becomes paramount.

In this letter, we introduce a novel method, which, to the best of our knowledge, is the first to counter SCOAP-based cluster detection by focusing on designed HTs. Our proposed

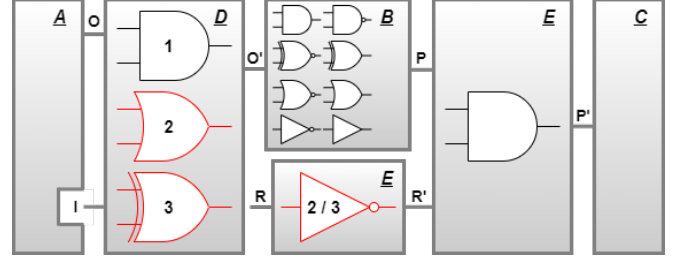


Fig. 1: The Structure of Proposed Method

approach involves the insertion of two correlated signals at specific locations within the HT. These insertion structures effectively reduce the extreme SCOAP values associated with HT signals, enabling them to remain inconspicuous within the genuine circuit. Notably, our insertion structures are designed to be simple, ensuring easy concealment within the circuit. Additionally, the non-uniqueness of the output results provides designers with greater flexibility. Furthermore, we guarantee that the trigger rate after reduction will be either lower or equal to the trigger rate before reduction, without imposing any constraints on the HT structure.

We conducted experiments to validate the effectiveness of our proposed method. The results clearly demonstrate that our method significantly reduces SCOAP values, irrespective of the structural restrictions of the HT. We also highlight the poor performance of SCOAP-based detectors, exhibiting high false positive rate (FPR) and false negative rate (FNR) when applied to the benchmarks after the reduction process. Additionally, we employ the commercial tool TetraMAX to verify that our proposed method does not introduce any redundancy. Overall, our method offers a promising solution for reducing SCOAP values in HT signals and evading detection by SCOAP-based cluster detectors.

## II. BACKGROUND

### A. SCOAP Testability Measures and Cluster Detection

The Sandia Controllability/Observability Analysis Program (SCOAP) is a topology-based program specifically designed for analyzing the testability of digital circuits. Testability refers to the ease or difficulty of controlling and observing signals within these circuits. Controllability refers to the difficulty in setting a specific logic signal to either 0 or 1, while observability relates to the challenge of observing the impact of a particular logic signal at the output. According to the definition of SCOAP measures, signals that are difficult to

activate will also be challenging to control and observe, resulting in high controllability or observability values. SCOAP-based detection assumes that all signals of HT possess extreme values and aims to identify them. A SCOAP-based detection technique is first presented in COTD [7]. COTD utilizes the  $CCs = \sqrt{CC0^2 + CC1^2}$  and observability ( $CO$ ) as the two-axis for analysis, and then employs unsupervised clustering to identify the HT signals. Unlike other techniques, COTD does not require a golden circuit or test pattern as a reference. Moreover, COTD demonstrates a 0% FPR and 0% FNR on Trust-hub [1] [2] [3] benchmarks, indicating its high accuracy in detecting HTs. Lastly, COTD exhibits a linear time complexity, implying efficient computational performance.

### B. Hardware Trojan benchmarks generation

The first benchmarks were published on Trust-Hub [1] [2] [3], which provides commonly used HT benchmarks for training and testing purposes. However, it is important to note that these benchmarks exhibit a wide distribution of trigger probabilities. The COTD detection has been proven to have good resolution for detecting HT benchmarks triggered by ultra-low probabilities. In recent studies such as [8] and [6], researchers have proposed new methods for generating benchmarks. They claim that the HT benchmarks produced using their methods can successfully evade COTD detection. However, it is worth mentioning that these methods have their own limitations. For example, in [8], their benchmarks have a small payload and some trigger rates are not low enough. Additionally, in [6], their proposed method can only be used for circuits in which the trigger signals of HTs are constructed using only *AND* gates. Although the structure they proposed can reduce the SCOAP value, we conducted experiments and the results revealed that no such structure were found in genuine circuits, indicating that once the detector considers their provided structure as one of the detection features, it can efficiently identify signals of HT.

## III. PROPOSED METHOD

In this section, we address the real-world scenario where HTs are intentionally designed, aiming to evade and attack detection. To facilitate the analysis and validation of our proposed method, we will introduce specific notations. Additionally, we will provide an overview of the proposed method, outlining its key principles and steps. Subsequently, we will delve into the detailed steps involved, accompanied by a comprehensive proof to substantiate the effectiveness of our approach.

To begin, we introduce three notations that are relevant to the positive probability:

- $\mathbf{Pr}_o(\mathbf{x})$ : Represents the positive probability of signal  $x$  before SCOAP value reduction.
- $\mathbf{Pr}_n(\mathbf{x})$ : Represents the positive probability of signal  $x$  after SCOAP value reduction.
- $\mathbf{Pr}_g(\mathbf{x})$ : Represents the positive probability of signal  $x$  which  $Pr_o(x) = Pr_n(x)$ .

Because the proposed method are aimed at HTs, it has a greater impact on the probability of HTs, so it is mostly

used for  $Pr_o$  and  $Pr_n$  to HT signals. In comparison, because most HTs have little impact on general circuits, the positive probability of general circuits has not changed, so we use  $Pr_g$  on general circuit signals. Moreover, we have  $\mathbf{Ap}(\mathbf{x})$  to represents the set of primary input patterns that can make signal  $x$  active.

### A. Overview

The fundamental structure of the proposed method is illustrated in Figure 1. The process begins by identifying the target signal  $O$  for which we aim to reduce the associated SCOAP value. Subsequently, we determine the insertion location  $P$  that ensures  $Pr_n(\text{trigger}) \leq Pr_o(\text{trigger})$ .

Next, in accordance with the conditions outlined in the proposed method, we select the signals  $I$  and  $R$ . These signals will play crucial roles in the subsequent steps of the method. We then introduce the insertion areas  $D$  and  $E$ . Area  $D$  consists of signals  $I$  and  $O$  as inputs, with  $O'$  as the output. Similarly, area  $E$  comprises nets  $P$  and  $R$  as inputs, with  $P'$  as the output. Once the design of areas  $D$  and  $E$  is complete, we insert them into the corresponding positions as depicted in Figure 1. This involves replacing the signal  $O$  with  $O'$  and the signal  $P$  with  $P'$ .

After the insertion process, the original HT is divided into three areas: area  $A$ , area  $B$ , and area  $C$ . Among these areas, area  $B$  does not require any special design considerations or a detailed understanding of its structure. By following the proposed method and incorporating the insertion areas  $D$  and  $E$ , we can effectively reduce the SCOAP value associated with the target signal  $O$  and ensure that the upper bound of  $Pr_n(\text{trigger})$  is met.

### B. Step 1 : Choose the type of SCOAP value reducer

The formulas for calculating the SCOAP value, as utilized in the proposed method, are listed in Table I. The core concept of the SCOAP value reducer is to employ the *min* function within the *AND*, *OR*, and *XOR* gates. To mitigate the high testability value caused by the signal  $O$ , it is crucial to carefully select the appropriate reducer in area  $D$ .

The proposed method offers three types of reducers, each designed to address specific situations:

1) **Type one**: This type of reducer is used to reduce high  $CC0$  of signal  $O$ . It is constructed using an *AND* gate in area both  $D$  and  $E$ . In this type,  $R = R'$ .





2) **Type two**: This type of reducer is used to reduce high  $CC1$  of signal  $O$ . It is constructed using an *OR* gate in area  $D$ , *Not* gate connect with *AND* gate in area  $E$ .

3) **Type three**: This type of reducer is used to reduce both  $CC0$  and  $CC1$  of signal  $O$ . It is constructed using an *XOR* gate in area  $D$ , *Not* gate connect with *AND* gate in area  $E$ .

### C. step 2 : Choose the signal $P$

The proposed method proceeds to determine the insertion location  $P$  for area  $E$  to be inserted. To accomplish this, we employ a reverse-engineering process to analyze the relationship between trigger signals. We focus on identifying

TABLE I: The candidates of  $\Phi$ 

$\Phi$	CC0(c)	CC1(c)	Pr(c)
	$\min\{CC0(a), CC0(b)\} + 1$	$CC1(a) + CC1(b) + 1$	$Pr(a) * Pr(b)$
	$CC0(a) + CC0(b) + 1$	$\min\{CC1(a), CC1(b)\} + 1$	$Pr(a) + Pr(b) - Pr(a)(b)$
	$\min\{CC0(a) + CC0(b), CC1(a) + CC1(b)\} + 1$	$\min\{CC1(a) + CC0(b), CC0(a) + CC1(b)\} + 1$	$Pr(a) + Pr(b) - 2Pr(a)Pr(b)$
	$CC1(a) + 1$	$CC0(a) + 1$	$1 - Pr(a)$

signals that, when their probability decreases, result in a corresponding decrease in the active probability of the payload. These signals become potential candidates for the insertion location  $P$ . By identifying these signals and treating them as candidates for insertion, we create opportunities for a wider range of combinations and outcomes. While it may not always be possible to identify all the relationships among HT signals, we can certainly identify a candidate that serves as the output of the HT trigger.

#### D. Step 3 : Choose the signals $I$ and $R$

For the type one, two and three reducers, we choose a signal with low  $CC0$ , low  $CC1$  and both low  $CC0$  and  $CC1$  as net  $I$ , respectively. After determining the signal  $I$  based on the SCOAP reducer types, the selection of signal  $R$  follows specific rules. For type one reducer, we select the  $R$  which  $Ap(R) \subseteq Ap(I)$  and, for type two & three reducer, we select  $R$  which  $Ap(I) \subseteq Ap(R)$ . By utilizing the subset relationship in the selection process, we ensure the diversity of outcome combinations and guarantee that there will be at least one signal  $R$  that is identical to signal  $I$  itself, regardless of the type of selection.

Indeed, the proposed method introduces a side effect of increasing the size of the HT. However, designers can optimize this aspect by selecting other qualified signals  $I_{new}$  based on the inserted  $R$  during the selection process. By doing so, the generation of new gates can be reduced, providing designers with greater control over the size expansion of the HT.

#### E. The structure of area $B$

We must to demonstrate that regardless of the structure of area  $B$ , we still can ensure that  $Pr_n(trigger) \leq Pr_o(trigger)$ . Specifically examining circuit area  $B$ , when the signal traverse from  $O$  to  $P$  before reduction, it passes through  $n$  logic gates. We define one logic gate as one operator  $\Phi$  and the candidates of  $\Phi$  are listed in Table I. The other input net of each  $\Phi_i$  as  $N_i$ , we can establish a recurrence relation expressed as:

$$Pr_o(T_i) = \begin{cases} Pr_o(O), & i = 0 \\ \Phi_i(Pr_g(N_{i-1}), Pr_o(T_{i-1})), & 1 \leq i < n \\ Pr_o(P), & i = n \end{cases} \quad (1)$$

To ensure upper bound of  $Pr_n(P')$ , we have hypothesis that  $Pr_n(P') = Pr_n(P) * Pr_g(R') \leq Pr_o(P)$ . For type one reducer, it is trivial that hypothesis holds true. For type two and type three reducers, when  $i = 0$ , we have  $Pr_n(O') = Pr_g(O) + [1 - Pr_g(O)]Pr_g(I)$  for type two and  $Pr_n(O') = Pr_g(O) + [1 - 2Pr_g(O)]Pr_g(I)$  for type three. Both of them

can simplified to  $Pr_n(O') = Pr_g(O) + \rho$ , where  $Pr_g(I) \subseteq \rho$ . Next, applying the iteration operator  $\Phi$ , when the probabilities of inputs are  $Pr_n(T_k) = Pr_o(T_k) + \rho_k$  and  $Pr_g(N_k)$ , in all four candidates in  $\Phi$  listed in Table I, the probability of output can still be reduced to  $Pr_n(T_{k+1}) = Pr_o(T_{k+1}) + \rho_{k+1}$ . After  $n$  iteration, we can conclude that  $Pr_n(P) = Pr_o(P) + \rho_n$ . As  $Ap(\bar{R}) \cap Ap(I) = \emptyset$ , it follows the hypothesis. Therefore, for all three types of reducer, we do not need to consider the structure of area  $B$ .

## IV. EXPERIMENTS

### A. Experimental Setup

To demonstrate the feasibility and effectiveness of our proposed method, we applied it to four benchmarks available on Trust-hub:  $s15850-T100$ ,  $s35932-T100$ ,  $s38417-T100$ , and  $s38584-T100$ . The clustering detector COTD was able to perfectly distinguish HTs in  $15850-T100$ ,  $35932-T100$ , and  $38417-T100$  with 0% FPR and 0% FNR. However, when using the  $s38584-T100$  benchmark, COTD can only achieve a FPR of 20% and a FNR of 11%.

### B. Reduction Steps

Our approach begins by identifying a set of signal candidates, denoted as  $S$ , capable of accommodating an area of  $D$ . Additionally, we define the target signal set,  $Q$ , which encompasses the signals for which we aim to reduce the SCOAP value. To preserve the structure and function of the payload, we solely consider the input signals of the HT and the trigger signals as part of  $S$ . Since the input signals of the HT are always genuine, we exclusively select trigger and payload signals for  $Q$ .

To minimum the number of new gate generation, in this experiment, we use the same  $R$  for all reduce iteration, which means all  $I$  follows the rule  $Ap(R) \subseteq Ap(I)$  or  $Ap(I) \subseteq Ap(R)$ . Subsequently, we systematically apply SCOAP value reducers to all signals in  $S$  using a brute force approach. Next, we calculate the  $CC$  value for each signal in  $Q$  using the formula  $CC = \sqrt{CC0^2 + CC1^2}$ . The resulting  $CC$  values are aggregated as  $CC_{total}$ . Finally, we select the results with the smallest  $CC_{total}$  as the input for the subsequent iteration. This iterative process continues until all signals in  $S$  have undergone SCOAP value reduction.

### C. Result and Analysis

Table II displays the experimental results, with "avg.CC" and "avg.CO" columns providing insights into the average  $CC$  value and average  $CO$  value of the target signals  $Q$ , respectively. The data in the "avg. CC" columns show an

TABLE II: The experimental results of SCOAP reduction

Benchmarks	SCOAP reduction													COTD[7]				
	avg. CC (reducing rate %)					avg. CO (reducing rate%)		Area $E$				gate reduction			before		after	
	0	1	2	3	all	0	all	1	2	3	$\Delta$	before	after	$\nu$ (%)	FPR(%)	FNR(%)	FPR(%)	FNR(%)
s15850-T100	1617(0)	1017(37)	776(52)	693(58)	16(99)	1734	79(96)	3	4	5	60	173	63	64	0	0	15	100
s38417-T100	2013(0)	1669(18)	1343(34)	1226(40)	16(99)	5351	75(99)	2	3	6	36	64	29	55	0	0	18	100
s38584-T100	58(0)	41(30)	33(44)	27(54)	19(67)	137	70(49)	2	4	5	40	43	21	51	20	11	20	100
s35932-T100	1268(0)	1018(20)	925(28)	843(34)	25(98)	124	65(48)	2	5	6	60	94	34	64	0	0	53	11
Average	1239(0)	936(26)	769(40)	697(47)	19(91)	1837	72(73)	2.3	4	5.5	49	95.8	72.8	58.5	5	2.8	33.2	78

average reducing rate of 26.3%, 39.5%, and 46.5% for the first three iterations, and up to 91% when all candidate signals in  $S$  undergo reduction. Furthermore, the proposed method achieves an average reduction of 73% in the  $CO$  value after the completion of the reduction process. It is worth noting that benchmarks  $s15850 - T100$ ,  $s38417 - T100$  and  $s38417 - T100$  experience an impressive reduction of almost 99% in their  $CC$  values, while  $s38584 - T100$  shows a smaller reduction. This discrepancy can be attributed to the fact that the  $CC$  value in  $s38584 - T100$  is already quite low, thus limiting the extent of reduction. Moreover, the final  $CO$  values for all four benchmarks are the similar, with the differences in reducing rate stemming from the original values. In reality, the actual average  $CC$  and  $CO$  values of the four benchmarks after reduction are very similar to each other. These experimental results clearly indicates that the proposed method significantly reduces both  $CC$  and  $CO$  values simultaneously. Overall, for benchmarks with a significant gap between HTs and general signals, our method effectively reduces and narrows down the differences in  $CC$  and  $CO$  values.

Additionally, the column "Area  $E$ " represents the number of places where  $E$  can potentially be inserted for each iteration, and the column  $\Delta$  represents the possible outcomes after three iterations. Notably, the average of 49 combinations after three iterations is the minimum number of combinations if we consider the selection of  $I$  and  $R$ . This indicates the versatility and effectiveness of our method in producing different outcomes. Furthermore, the results of the new gate generation, whether choosing different or the same  $R$  for all iterations, are presented in the "gate reduction" column, with the corresponding reduction ratio displayed in column  $\nu$ . The analysis reveals that approximately 58.5% of the new gates experience reduction.

Moreover, when evaluating COTD using the reduced benchmarks, the results show an average FPR of 33.3% and a FNR of 78%. For  $s15850 - T100$ ,  $s38417 - T100$ , and  $s38584 - T100$ , the proposed method can perfectly conceal the HT within genuine signals. However, for  $s35932 - T100$ , the proposed method can only provide an 11% FNR. This is because the  $CO$  value is calculated from the direction of output to input, meaning that the  $CO$  value of the HT is based on the signal to which its output is connected. In the case of  $s35932 - T100$ , the HT is connected to a signal with a high  $CO$  value, this signals will also be mistaken for an HT. Nevertheless, the proposed method can still provide a 53% FPR value, forcing the detector to exert significant efforts to scrutinize all suspicious candidates and verify the presence of the real HT. The elevated values of both FPR and FNR indicate that our proposed method renders the cluster detector ineffective. Lastly, we compared the fault lists

generated by TetraMAX and found that none of the four reduced benchmarks produced any new redundancies.

## V. CONCLUSION

In this letter, we present the first-ever method to effectively evade SCOAP-based cluster detection for both designed and detected HTs. Our approach introduces no structural limitations and leverages simple circuit structures commonly found in general circuits for insertion. The output results after reduction are diverse, providing HT designers with a broader range of output choices and enhancing the stealthiness of HTs. The experimental results demonstrate that proposed SCOAP value reducers significantly decrease SCOAP values while ensuring an upper bound on the trigger probability for the payload. The reductions achieved with the  $CC$  and  $CO$  values average an impressive 91% and 73%, respectively. Moreover, when evaluating the circuits with the new results, SCOAP-based cluster detection exhibits high FPR and FNR, indicating poor performance and rendering the detection approach ineffective. This highlights the robustness of our method in evading detection and concealing the presence of HTs within the normal signal. To further validate the integrity and functionality of our circuit design, we employ TetraMAX testing, ensuring that our proposed method introduces no redundancies. This ensures the reliability and safety of the circuit, without compromising its performance.

## REFERENCES

- [1] Trust-hub. Available on-line: <https://www.trust-hub.org>, 2016.
- [2] B. Shakyia et al., "Benchmarking of HTs and Maliciously Affected Circuits", Journal of Hardware and Systems Security (HaSS), April 2017.
- [3] H. Salmami, M. Tehranipoor, and R. Karri, "On Design vulnerability analysis and trust benchmark development", IEEE Int. Conference on Computer Design (ICCD), 2013.
- [4] M. Tehranipoor and F. Koushanfar, "A Survey of Hardware Trojan Taxonomy and Detection," in IEEE Design & Test of Computers, 2010.
- [5] R. S. Chakraborty, S. Narasimhan and S. Bhunia, "Hardware Trojan: Threats and emerging solutions," 2009 IEEE International High Level Design Validation and Test Workshop, 2009.
- [6] C.-W. Chen et al., "A Hardware Trojan Insertion Framework against Gate-Level Netlist Structural Feature-based and SCOAP-based Detection," 2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS), Fukuoka, Japan, 2022.
- [7] H. Salmami, "COTD: Reference-Free Hardware Trojan Detection and Recovery Based on Controllability and Observability in Gate-Level Netlist," in IEEE Transactions on Information Forensics and Security, 2017.
- [8] J. Cruz et al., "An automated configurable Trojan insertion framework for dynamic trust benchmarks," 2018 Design, Automation Test in Europe Conference & Exhibition (DATE), 2018.
- [9] P.-Y. Lo et al., "Semi-supervised Trojan Nets Classification Using Anomaly Detection Based on SCOAP Features," 2022 IEEE International Symposium on Circuits and Systems (ISCAS), 2022.
- [10] L. H. Goldstein and E. L. Thigpen, "SCOAP: Sandia Controllability/Observability Analysis Program," 17th Design Automation Conference, 1980.