A Multi-Dimensional Hardware Trojan Design Platform to Enhance Hardware Security

Nilanjana Das, *IEEE, Member*, Mattis Hasler, *IEEE, Member*, Friedrich Pauls, Sebastian Haas

***Abstract*—This paper proposes a novel kind of HT design named multi-dimensional hardware Trojans (MDHTs) and develops a method to generate configurable MDHT benchmark platform. The proposed MDHT circuits include multiple net(s) as trigger signals from each of the rarely activated, highly activated, and partially activated categories to increase MDHT's adverse effects. The generated MDHT-infected circuits are tested by an unsupervised machine learning based HT detection technique-Controllability and Observability for HT Detection (COTD). Experimentation on ISCAS benchmarks ensures that the detection method is unable to detect the developed MDHT circuits as the nets belong to higher and partial activities are creating at least 50% and at most 80% false negative rate which validates the MDHT insertion framework in addition to the available HT benchmarks.**

***Index terms* – Multi-dimensional HT (MDHT), benchmark, transition probability, controllability, observability, k-means clustering.**

I. Introduction

The research community has made immense progress in the development of efficient countermeasures against existing types of HTs, these are defeated by sophisticated HTs created afterward. Static benchmark suite [1] is a great contribution to the standardization of HT testing, however, as the Trojan location and trigger conditions are static, detection techniques can be optimized to target these HTs rather than generic HTs. Moreover, new types of HTs cannot be updated into these HTs in a timely manner. To address these limitations, a tool for generating dynamic HT circuits was presented in [2], [3]. These automatic HT insertion frameworks dynamically insert HTs into gate-level designs based on rare internal nodes identified based on functional simulation (FS) and transition probability (TP) calculation respectively. Although the main objective of the HT research community is to propose strong detection strategies, they are unable to make significant progress due to a lack of heterogeneous HT benchmarks. Rigorous research work is therefore required on possible HT designs to make the detection strategies more powerful.

To improve these drawbacks, several automatic HT insertion frameworks were presented as follows [1]–[3]. It is proved in [4] that when the value of TP is low for a net, the testability (CC) becomes imbalanced and this low TP value was used to generate stealthy HT benchmarks. However, in [5] it is stated that a net with low TP can have balanced CC and provide detection approaches accordingly. Also, all works mentioned above as well as [6]–[8] consider nets with low switching activity or rare nets to create a trigger using a single trigger signal (TS) for their HT design but new scenarios can arise after consideration of multiple TSs in an HT circuit are still missing. Hence, these works presents following three drawbacks: (i) the FS is considered for rare net selection [2] does not provide a realistic reflection of net switching activities when testing patterns are insufficient; (ii) consideration of TP for activity calculation provides better view of net's activity than FS [3]. However, CC analysis of each net is not performed for activity calculation in this case; and (iii) for all the works, only the rare nets are selected for Trojan circuit generation. Consideration of multiple TSs with the combination of rarely, partially, and highly activated nets are missing in these literature.

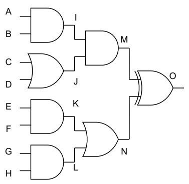
To fill the above-mentioned limitations in HT design, in this paper, we propose a novel method to generate Multi-Dimensional HTs (MDHTs) using an HT generation platform based on both TP and CC [9] to get an appropriate result of the activation of each net. To the best of our knowledge there are no prior work considered both TP and CC to calculate the activity of a net. In summary, our contribution in this work is described as follows. (i) A new parameter activity (*AC*) is defined considering both the TP and the CC metrics (0-controllability, 1-controllability) to compute the activity (*AC*) of each net in the netlist. (ii) The MDHT considers more than one TSs in the HT design which also raise the point that TSs can be partially as well as highly activated apart from rarely activated nets. (iii) The strength of MDHTs over the existing HT detection procedures are demonstrated analytically. (iv) A heuristic model to generate random-MDHT benchmarks is developed for a given core using the parameter activity (*AC*). (v) The efficiency of developed MDHT benchmarks is tested using the unsupervised k-means clustering based detection technique COTD [10]. Experimental results indicate that COTD is unable to detect a proper subset of TSs due to high false negative (FN) rate on all generated MDHT benchmarks.

II. Parameters to Determine Switching Activity

In this section, an overview of existing kinds of parameters to determine the switching activity of a net is described. FS uses statistical analysis to estimate the switching activity of internal nets [2]. The TP is defined as the estimated time required to generate a transition on a net by performing the geometric distribution [3].

The Sandia Controllability/Observability Program (SCOAP) [9] is one of the most popular testability programs that measures CC of each net(s) in a circuit logic based on following numerical values: combinational 0-controllability of a net s (CC0(s)) which indicates the difficulty of setting the signal to 0, combinational 1-controllability of a net s (CC1(s)) which indicates the difficulty of setting the signal to 1, and combinational observability of net s (CO(s)) [9]. From [10], HT signals should have poor testability, which implies they have high CC values or high CO values.

The authors are with the Barkhausen Institut, 01067 Dresden, Germany (e-mail: firstname.lastname@barkhauseninstitut.org).



Net	TP	CC0, CC1	*CC*	|CC0-CC1|	*DCC*	*AC*
I	0.1875	2, 3	3.6055	1	0.33	0.1256
J	0.1875	3, 2	3.6055	1	0.33	0.1256
K	0.1875	2, 3	3.6055	1	0.33	0.1256
L	0.1875	2, 3	3.6055	1	0.33	0.1256
M	0.1523	3, 6	6.7082	3	1	0
N	0.2460	5, 4	6.4031	1	0.33	0.1649

Nets in increasing order of *AC* = {M, I, J, K, L, N}

(a) Sample circuit (b) Values of TP, *DCC*. and *AC* of each net

Fig. 1: Illustration of TP, *DCC*, and *AC*.

When TP and CC values are considered individually for rare net selection, they can provide different activation behaviors of that net. To illustrate the concept of *AC*, a sample circuit is shown in Fig. 1a and the TP, CC0, CC1, CC, and *DCC* values of each internal net is presented in Fig. 1b. For example, in Fig. 1b, net N has a high TP value (0.2460) which suggests it is highly active [3] and not suspicious for HT insertion. Also, N has a high CC value (6.4031) which denotes poor testability and it is rarely activated and highly suspicious for HT insertion [10]. These two situations contradict the fact that low TP denotes rare activated net. To resolve this confusion, we have proposed a new parameter *AC* described in Eq. (1b) which provides a better view of switching activity of a net as it considers both the TP and CC.

A net is rare if it has low TP and high $|CC0-CC1|$[1]. At first we define a new parameter called Deviation of Testability Values (*DCC*) for a net s using Eq. (1a) where N is the set of all nets in the core. Then *DCC* and TP both parameters are used to compute the *AC*. A net is rarely active if it has low TP and high *DCC*, hence low *AC*. Similarly, a net is highly active if it has high TP and low *DCC*, hence high *AC*. Note that, range of *AC* is $[0,1]$ as both TP and *DCC* are in range of $[0,1]$. We measure TP, CC0, CC1, and *DCC* for each net and define activity (*AC*) of a net using Eq. (1b) describe as follows.

$$DCC_s = \frac{|CC0-CC1|_s}{max_{s\in N}|CC0-CC1|} \tag{1a}$$

$$AC = \text{TP}\cdot(1-DCC) \tag{1b}$$

The CC value of a net is defined by $CC = \sqrt{CC0^2+CC1^2}$ [4]. Assuming probability of random inputs (Pr(1) = Pr(0) = 0.5) are applied to the input nets of Fig. 1a, the TP of each net can be calculated based on the logic function of each gate. According to the truth table of an AND gate [3], Pr(1) and Pr(0) of net K is 1/4 and 3/4, so the TP of net K is (Pr(0)) × (Pr(1)) = 0.1875. For net K the CC0(K) is (min(CC0(E), CC0(F)) + 1) = min(1, 1) + 1 = 2. The CC1(K) is (CC1(E) + CC1(F) +1) = (1+1+1) = 3. Net M has least *AC* as it has the lowest TP and the highest *DCC* values. From Fig. 1b, net N has high TP value (0.2460) which suggests it is highly activated [3] but high CC value (6.4031) suggests it is rarely activated [10]. However, the *AC* value of the net N gives a better estimation and shows that N is actually a highly activated net instead of rarely activated net though it has high CC (poor testability) value. By this way, our proposed parameter *AC* helps to better deduce the activation of a net rather than considering TP and CC individually.

III. Concept and Strength of MDHT

Concept of MDHT: Let, T be the set of TS(s) involved in an HT circuit where $T = \{t_1, t_2, ..., t_d\}$. Each net in T is categorized into following three sets based on *AC*. A_L: The TSs with a low *AC* belong to set A_L, i.e. $t \in A_L \Rightarrow$ TP $\rightarrow 0$, $DCC \rightarrow 1$, and $AC \rightarrow 0$. A_H: The TSs with a high *AC* belong to set A_H, i.e., $t \in A_H \Rightarrow$ TP $\rightarrow 1$, $DCC \rightarrow 0$, and $AC \rightarrow 1$. A_M: The TSs belong to set A_M if they have *AC* higher than the TS $\in A_L$ and lower than the TS$\in A_H$. Therefore, if $t_1 \in A_L$, $t_2 \in A_H$, and $t_3 \in A_M$ then $AC(t_1) << AC(t_3) << AC(t_2)$.

Definition 1: An HT circuit is an MDHT if $|T| \geq 3$. In this case, at least three TSs are involved in that HT circuit to activate the payload.

Definition 2: In an d-dimensional HT, to propagate the malicious outcome instead of the original outcome, all the TSs must be in active state simultaneously at a specific timestamp, that is $\forall i \in [1,d]$, state of $t_i = 1$.

Strength of MDHT: The strength of MDHT depends on the number as well as the type of nets involved in the HT. The necessary condition of an MDHT remains undetected by the existing detection algorithms is $|A_L| \geq 1$, $|A_H| \geq 1$, and $|A_P| \geq 1$. The following three lemmas describe the functionalities of each type of net to evade the detection algorithms.

Lemma 1: **The detection algorithms with prime concern to find out the rare nodes in an HT circuit cannot detect the MDHT if $|A_H| \geq 1$.** Let in an MDHT, $|T| \geq 3$ where $|A_H| \geq 1$ and $t \in A_H$. Hence, t will activate frequently. Therefore, t cannot be suspected by the detection algorithms with the prime concern being to detect rarely activated nets. Again, from **Definition 2**, the malicious outcome is observed when all the nets in T are activated simultaneously. If T^* is the set of suspected nets by a detection algorithm, then $t \notin \text{T}^*$ because $t \in A_H$. Therefore, T^* does not contain all the nets from T (i.e., $t \in A_H \implies t \notin T^*$), and it results MDHT remain undetected by the aforementioned detection algorithm.

Lemma 2: **MDHT will not exhibit malicious behavior frequently if $|A_L| \geq 1$.** Let in an MDHT, $|T| \geq 3$ where $|A_L| \geq 1$ and $t \in A_L$. Hence, t will remain inactive most of the time. Therefore, t will be suspected by the detection algorithms with the prime concern being to detect rarely activated nets. However, from Lemma 1, if $|A_H| \geq 1$, MDHT will remain undetected. Again, from **Definition 3**, the malicious outcome is observed when all the nets in T are activated simultaneously. As t activate rarely, the MDHT will not show malicious behavior frequently, and remain undetected during run-time.

Lemma 3: **Dimension of the MDHT is unrevealed if $|A_M| \geq 1$.** The nets belong to A_M in an HT circuit creating an impression that they are part of the original circuit. In other words, the detection strategies cannot identify the exact value of d need to be suspected before identifying the MDHT.

When the TSs are combination of A_L, A_H and A_M, then detection of A_L cannot guarantee of involved TSs belongs to A_H and A_M. The list of captured wires is T^*. As A_H is not in the suspected list then $t \in A_H$ denotes $t \notin T^*$. Therefore, T should contain combination of nets that are belongs to

[1] From [4], a net is rare if distance of (0,0) and $(\frac{CC0}{CC1}, \frac{CC1}{CC0})$ (=d') is high. It can be verified that d' is high if $|CC0-CC1|$ is also high.

Algorithm 1 MDHT Insertion Algorithm

```
Require: Set of nets in netlist.v ($N$), $\phi$, $\theta_l$, $\theta_h$.
Ensure: Ensure netlist.v with inserted MDHT of dimension $d$.
Compute $d = \lfloor \frac{|N|}{\phi} \rfloor$
if $d < 3$ then
    Exit.
else
    for all net $\in N$ do
        Compute TP, $DCC$, $AC$.
    Sort $N$ in increasing order of $AC$. Divide all nets in three classes: $R$, $P$, and $H$ considering $\theta_l$ and $\theta_h$.
    if $d = 3$ then
        Set $l = m = h = 1$.
    else
        $\overline{d} = \lfloor \frac{d}{3} \rfloor$, randomly assign $l$, $m$ and $h$ from the ranges [1, $\overline{d} - 1$], [1, $\overline{d}$] and [1, $\overline{d} - 1$].
    Randomly select $l$, $m$ and $h$ number of nets as TSs from $R$, $P$, and $H$ respectively. Flag=False.
    while Flag=False do
        Randomly select $(l+m+h)$ nets in topological order from the output of circuit. Insert $(l+m+h)$ number of 2 to 1 multiplexer with selected TSs as select line. The inputs of 2 to 1 multiplexers: a gate from original circuit, a replacement gate of selected gate as original input.
        if Inserted MDHT is valid then
            Flag=True. Exit().
```

TABLE I: Different scenarios of Payload insertion in MDHT

Trigger signal type	Gate					
	AND	OR	NAND	NOR	XOR	XNOR
R	NAND NOR, XOR	NAND NOR, XOR	XNOR OR, AND	XNOR, OR, AND	XNOR AND, OR	XOR NOR, NAND
P	NOR	NAND	OR	AND	-	-
H	-	XOR	-	XNOR	OR	NOR

A_L, A_H and A_M to avoid the detection. Hence, these three lemmas manifest as follows. (i) The A_H is responsible to evade detection algorithms. (ii) The A_L is responsible for the rare explicit behavior of the Trojan circuit. (iii) The A_M is responsible for creating confusion to the detection algorithm about the dimension of MDHT. Hence, it can be concluded that $|A_L| \geq 1$, $|A_M| \geq 1$, and $|A_H| \geq 1$ is sufficient in an MDHT to evade the detection algorithms.

IV. Proposed MDHT Generation Platform

A key aspect of MDHT design is that the HT circuit should be small enough compared to the golden netlist so that the area overhead (OA) and power overhead (OP) doesn't show any suspicion. An automatic d generation platform is designed to incur negligible OA, OP with respect to the original netlist.v due to MDHT insertion. We define a circuit-dependent input parameter ϕ as the number of nets required corresponding to one TS, i.e., $d = \lfloor \frac{|N|}{\phi} \rfloor$ where N is set of nets in netlist.v. Algorithm 1 describes the proposed MDHT generation platform in a given IP core.

MDHT trigger signals (TSs) insertion: Algorithm 1 inserts MDHT in an IP core if $d \geq 3$. At first, TP, DCC, and AC values for each net are calculated. Next, N is sorted in ascending order of AC and the sets of rarely active nets (R), highly active nets (H), and partially active nets (P) are computed considering input parameters θ_l and θ_h as follows. For a net s, if AC(s) $< \theta_l$ then $s \in R$ and if $AC(s) > \theta_h$ then $s \in H$; and $P = N - (R \cup H)$. Let, l, m, and h are the cardinality of A_L, A_M, and A_H. To insert the MDHT in an IP core l, m, and h number of nets are randomly selected from R, P, and H as TSs. The values of l, m, and h are determined in such a way that $(l+m+h) \leq d$ and $m \geq l, h$, which create confusion about the dimension of inserted MDHT.

TABLE II: Determination of MDHT dimension and other parameters based on ϕ.

Benchmarks	Area (μm^2)	Power (μW)	ϕ	d	OA(%)	OP(%)	$\|R\|$, $\|P\|$, $\|H\|$	l, m, h
C499	690.899	25.36	**60**	3	0.92	0.98	17, 170, 28	1, 1, 1
C1908	733.109	27.2	**600**	3	0.78	0.89	39, 411, 62	1, 1, 1
C2670	1647.049	61.8	**600**	4	0.83	0.82	76, 815, 132	1, 2, 1
C3540	1896.684	75.0	**800**	4	0.76	0.78	81, 865, 147	1, 2, 1
C5315	2693.474	103.0	**1000**	5	0.69	0.73	124, 1386, 228	2, 2, 1
C6288	3982.314	223.0	**1000**	6	0.72	0.65	173, 1884, 328	2, 2, 1
b17	24800.814	1020.39	**3200**	9	0.35	0.28	1068, 22583, 4199	2, 3, 1

Payload insertion: In this work, we assume that all these TSs act as select lines of 2 to 1 multiplexers which are the parts of the payload circuit. A set of $(l+m+h)$ nets are randomly selected in topological order from the output of the core. The inputs of 2 to 1 multiplexers are the selected gates and a replacement gates from the core (original) as shown in Table I for R, P, and H, respectively. The TSs belongs to R will be closest one and and TSs from H will be farthest one from the original output of the core respectively. If the inserted MDHT is not valid according to [2], it is removed from the core and a new set of $(l+m+h)$ nets are selected for MDHT insertion.

V. Experimental Results

To measure the efficiency of proposed MDHT benchmark generation algorithm, we have considered ISCAS and ITC-99 benchmarks to demonstrate the effectiveness of proposed MDHT model. The platform supports reading gate-level netlist designs using the Cadence Genus (TM) Synthesis Solution with 45 nm CMOS technology. The COTD flow is applied to our experiments, which is based on unsupervised k-means clustering approach for Trojan detection. First, the TP, CC and CO values of all signals are computed using an open-source tool [11] with python 3.6, and the signals in the gate-level netlist can be classified using the k-means clustering algorithm. Algorithm 1 executes until a valid MDHT is inserted in the circuit and each simulation is performed 12 times.

Table II reports the area and power overhead of considered benchmarks after insertion of MDHT circuit, considered dimension (d) of inserted MDHT, cardinalities of R, P, and H, and number of TS(s) of each type. At first, random values of ϕ is considered from the range $[60, 1200]$ to compute value(s) of d. For a circuit, only those ϕ values are considered for which the OA and OP are less than 1% compared to the original one. Among these valid set of ϕ values, the least ϕ value is considered which results the highest value of d and the MDHT is more stealthy. The values of θ_l are θ_h are considered in percentage from the range $[6, 10]$ and $[10, 15]$ of total nets in the circuit. The sets R, P, H are determined based on selected θ_l are θ_h values. The number of nets belong to the sets R, H, and P are reported. The values of l, m, and h are determined based on selected value of d. Note that the value of $(l+h+m)$

TABLE III: Simulation Results of COTD-Based HT Detection On MDHT-inserted Benchmarks

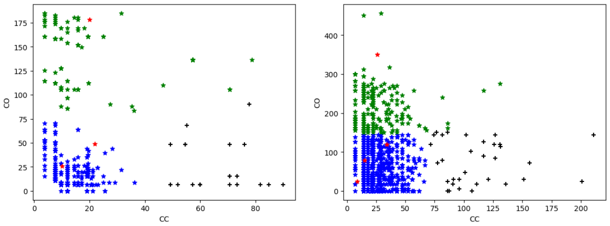
Benchmark	Centroid			No. of points			Location of TSs			TSs in cluster			FN	FP	(Normal/HT)		
	C_{NO}	C_{HC}	C_{HO}	C_{NO}	C_{HC}	C_{HO}	A_L	A_M	A_H	A_L	A_M	A_H	FNR(%)	FPR(%)	C_{NO}	C_{HC}	C_{HO}
C499	(8.54, 5.41)	(30.64, 12.77)	(9.17, 42.08)	105	46	90	(4, 55)	(11, 16)	(6, 7)	C_{HO}	C_{NO}	C_{NO}	13(50)	123(57)	92/13	40/6	83/7
C1908	(12.28, 17.52)	(65.01, 24.20)	(17.52, 140.54)	395	39	104	(20, 178)	(22, 49)	(10, 26)	C_{HO}	C_{NO}	C_{NO}	14(53)	131(25)	381/ 14	35/4	96/8
C2670	(11.93, 13.38)	(168.5, 62.79)	(13.28, 249.53)	889	26	146	(35, 319)	(49, 86), (91, 23)	(10, 37)	C_{HO}	C_{NO},C_{NO}	C_{NO}	20(52)	154(15)	869/20	21/5	133/13
C3540	(26.69, 51.95)	(110.54, 72.35)	(24.68, 211.36)	863	42	226	(26, 35)	(15, 78), (34, 120)	(9, 25)	C_{HO}	C_{NO}, C_{NO}	C_{NO}	19(50)	249(22)	844/19	33/9	216/10
C5315	(18.14, 31.68)	(62.34, 17.91))	(18.07, 168.26)	1587	121	135	(74, 9), (26, 230)	(24, 49), (36, 52)	(9, 22)	C_{HC}, C_{HO}	C_{NO}, C_{NO}	C_{NO}	48(80)	244(13)	1539/48	115/6	129/6
C6288	(64.08, 115.85)	(120.88, 117.45)	(77.52, 308.29)	1427	848	170	(190, 102), (75, 420)	(52, 92), (75, 80)	(40, 180)	C_{HC}, C_{HO}	C_{NO}, C_{NO}	C_{NO}	35(58)	993(41)	1392/35	828/20	165/5
b17	(130.17, 1180.22)	(470.25,1110.28)	(54.21,2234.54)	15784	10182	1958	(10, 2871), (600, 284)	(10, 102), (35, 1200), (78, 1136)	(105, 1400)	C_{HC},C_{HO}	C_{NO},C_{NO}	C_{NO}	54(75)	12122(44)	15730/54	10170/12	1952/6

is not always equals to d, e.g., for C6288 value of d is 6, but value of $(l+h+m)$ is 5.

For each circuit, all the nets are classified into three categories using the COTD detection method based on unsupervised k-means clustering approach with k=3. Three clusters are formed. C_{NO} (marked in blue)- the nets in this cluster has low CC and CO values, C_{HC} (marked in black)- the nets in this cluster have high CC values and low CO values, and C_{HO} (marked in green)- The nets in this cluster have low CC values and high CO values. Table III reports the centroids of each of these clusters and number of nets belongs to these clusters. Note that, these clusters also include nets from the inserted MDHT circuit. Table III presents location of the $(l+m+h)$ number of nets selected as TSs [Table II] and in which cluster(s) they belong. It is observed in all cases that TSs from A_L (i.e., TSs with low value of AC) are always belong to C_{HC} or C_{HO}. TSs from A_M and A_H (i.e., TSs with medium and high values of AC) are always belong to C_{NO}. As COTD does not suspect nets belongs to cluster C$_{NO}$, these nets are not detected as TSs. Therefore, TSs with partial or high AC remain undetected using COTD, and MDHT successfully evades COTD. These results are described in Fig. 2 where the inserted TSs are marked in red. Table III shows the false negative (FN), false negative ratio in percentage (FNR), false positive (FP), and false positive ratio in percentage (FPR) for inserted MDHT circuits [3]. It is observed that COTD generates higher values of FNR and FPR, which indicates that MDHT is able to outrun the COTD. Table III reports the number of nets from original circuit and number of nets from MDHT circuit for each cluster. In general, the detection outcomes demonstrate that *this platform can produce failed test circumstances with a high FNR for COTD detection on almost all created MDHT-infected circuits. The same is true for detection approaches presented in [4], [5] as they considered only rare nets with imbalanced and balanced CC, respectively.*

VI. Conclusion

In this paper, we propose a novel method to generate MDHTs using a HT generation platform by computing the activity of each net using both the transition probability and testability parameters. We identify the rarely activated, mostly activated and partially activated nets in a circuit to target for MDHT insertion. The platform has been tested to generate HT-infected circuits from ISCAS-85 benchmarks and evaluated by the COTD detection technique. Simulation results ensure that the rare nets can be detected by the COTD but the partially and mostly active nets used as trigger signals cannot be detected by the COTD. Moreover, the inserted MDHT results 50% to 80% false negative rate in all cases as well. Hence, it is near to impossible to detect MDHT with COTD or other HT detection algorithms looking for rare nets only as ensured by the obtained results.



(a) C1908

(b) C3540

Fig. 2: K-means Clustering examples on ISCAS-85 benchmarks using COTD.

References

[1] Bicky Shakya, Tony He, Hassan Salmani, Domenic Forte, Swarup Bhunia, and Mark Tehranipoor. Benchmarking of hardware trojans and maliciously affected circuits. *Journal of Hardware and Systems Security*, 1:85–102, 2017.

[2] Jonathan Cruz, Yuanwen Huang, Prabhat Mishra, and Swarup Bhunia. An automated configurable trojan insertion framework for dynamic trust benchmarks. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1598–1603. IEEE, 2018.

[3] Shichao Yu, Weiqiang Liu, and Maire O'Neill. An improved automatic hardware trojan generation platform. In *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 302–307. IEEE, 2019.

[4] Yu Su, Haihua Shen, Renjie Lu, and Yunying Ye. A stealthy hardware trojan design and corresponding detection method. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–6. IEEE, 2021.

[5] Wei-Ting Hsu, Pei-Yu Lo, Chi-Wei Chen, Chin-Wei Tien, and Sy-Yen Kuo. Hardware trojan detection method against balanced controllability trigger design. *IEEE Embedded Systems Letters*, 2023.

[6] Farinaz Koushanfar Mohammad Tehranipoor, Ramesh Karri and Miodrag Potkonjak. Trusthub." [online]. available: http://trust-hub.org.

[7] Jie Zhang, Feng Yuan, Lingxiao Wei, Zelong Sun, and Qiang Xu. Veritrust: Verification for hardware trust. In *Proceedings of the 50th Annual Design Automation Conference*, pages 1–8, 2013.

[8] Adam Waksman, Matthew Suozzo, and Simha Sethumadhavan. Fanci: identification of stealthy malicious logic using boolean functional analysis. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 697–708, 2013.

[9] Lawrence H Goldstein and Evelyn L Thigpen. Scoap: Sandia controllability/observability analysis program. In *Proceedings of the 17th Design Automation Conference*, pages 190–196, 1980.

[10] Hassan 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*, 12(2):338–350, 2016.

[11] Joseph Sweeney, Ruben Purdy, Ronald D Blanton, and Lawrence Pileggi. Circuitgraph: A python package for boolean circuits. *Journal of Open Source Software*, 5(56):2646, 2020.