TAAL: Tampering Attack on Any Key-based Logic Locked Circuits

Ayush Jain, Student Member, IEEE, Ziqi Zhou, Student Member, IEEE and Ujjwal Guin, Member, IEEE

Abstract—Due to the globalization of semiconductor manufacturing and test processes, the system-on-a-chip (SoC) designers no longer design the complete SoC and manufacture chips on their own. This outsourcing of design and manufacturing of Integrated Circuits (ICs) has resulted in a number of threats, such as overproduction of ICs, sale of out-of-specification/rejected ICs, and piracy of Intellectual Properties (IPs). Logic locking has emerged as a promising defense strategy against the aforementioned threats. However, various attacks pertaining to the extraction of secret keys have undermined the security of logic locking techniques. Over the years, researchers have proposed different techniques to prevent the existing attacks. In this paper, we propose a novel attack which can break any logic locking techniques that rely on stored secret key. This proposed TAAL attack is based on implanting a hardware Trojan in the netlist, which leaks the secret key to an adversary once activated. As an untrusted foundry has the capability to extract the netlist of a design from the layout/mask information, it is feasible for a malicious foundry to implement such a hardware Trojan. All of the three types of TAAL attacks can be used for extracting secret keys. We have introduced the models for both the combinational and sequential hardware Trojans that evade manufacturing tests as well. An adversary only needs to choose one hardware Trojan out of a large set of all possible Trojans to launch the TAAL attack.

Index Terms—Logic locking, IP Piracy, IC Overproduction, Hardware Trojans, Tampering

I. INTRODUCTION

The continuous addition of new functionality in a system-on-a-chip (SoC) has enforced designers to adopt newer and lower technology nodes to manufacture chips primarily to reduce overall area and the resultant cost of a chip. Building and maintaining such a fabrication plant (foundry) requires a multi-billion dollar investment [1]. As a result, the semiconductor industry has moved towards horizontal integration, where a SoC designer acquires intellectual properties (IPs) from many different vendors and sends the design to a foundry for manufacturing, which is generally located offshore.

The hardware layers that were assumed to be trusted, are no longer true with the outsourcing of IC fabrication in a globalized and distributed design flow including multiple entities. Third party IPs, fabrication, and test facilities of chips represent security threat to the current horizontal integration of the production. The security threats posed by these entities include – (i) overproduction of ICs [2, 3], where an untrusted foundry fabricates more chips without the consent of the SoC designer in order to generate revenue by selling them in the market, (ii) sale of out-of-specification/rejected ICs [8, 9], and (iii) IP piracy [10, 11], where an entity in supply chain can use, modify and/or sell functional IPs illegally. Over the years, researchers have proposed different techniques to prevent the aforementioned attacks and they are IC metering [2, 3, 5, 14], logic locking [2, 8, 15], hardware watermarking [10, 18], and split manufacturing [19, 20].

Logic locking has emerged as one of the promising technique and gained significant attention from the researchers to address the different threats emerged from untrusted manufacturing. In logic locking, the netlist of a circuit is locked in such a way that it produces incorrect results unless it is programmed with a secret key. The locks are generally inserted in the netlist using XOR gates. Traditional logic locking methods involved selection of key gate location as random selection [2], [9], [21], strong interference-based selection [15], and fault-analysis based selection [22]. Over the years, researchers have also proposed different attacks to extract the secret key and thus undermine the locking mechanisms. Boolean Satisfiability (SAT)-based attacks have demonstrated effective ways of extracting the secret keys. Countermeasures are also proposed so that SAT-based attacks become infeasible.

In this paper, we show that any locked circuits where the secret key is stored in an on-chip memory, even if tamper-proof, can be broken. We present this attack as TAAL: Tampering Attack on Any Locked circuits that use key-based logic locking. This attack can defeat the security measures provided from any existing logic locking methods. We believe that we are the first to show that any key-based logic locked circuits can be exploited by inserting a hardware Trojan in the netlist. The contributions of this paper are described as follows:

• We propose a novel attack based on malicious modifications of the netlist to target any key-based locked circuit. The attack approach is to tamper the locked netlist in order to extract the secret key information. Once the valid key information is extracted out from an activated IC, an untrusted foundry can unlock any number of chips, and sell overproduced and defective chips in the market. As this attack is applicable to any key-based locked circuits, an adversary can undermine any secure solutions proposed so far to prevent overproduction, sourcing defective/out-of-spec chips, and IP piracy.
• We present three types of TAAL attacks that extract the secret key differently using hardware Trojans placed at different locations in the netlist. T1 type TAAL attack directly leaks the secret key to the primary output once
the Trojan is activated (see Figure 2(a)). On the other hand, T2 type TAAL and T3 type TAAL attacks rely on the activation and propagation of the secret key to the primary output (see Figure 2(b) and Figure 3 respectively). The adversary has the freedom of choosing one of these three types of TAAL attacks implemented using combinational or sequential hardware Trojans.

- We present a model for the combinational hardware Trojan, which can be used to launch an attack. The adversary has the freedom of choosing the type of hardware Trojan, which can be designed in such a way that it evades manufacturing or production tests and remains undetected. We define this Trojan as Type-p Trojan, as it has \( p \) trigger inputs. These triggers can come from the primary inputs and/or internal nodes of a locked circuit. We show that a very large number of Trojans can be created and only one such a Trojan can be used by an adversary. It is practically possible for an SoC designer to detect all these feasible Trojans using logic tests.

- We also present a model for sequential Trojan, which is constructed using a combinational one. A state element (a counter) is added to a combinational Trojan so that it delivers the payload once it is triggered \( R \) times consecutively. Note that the trigger inputs for both the Trojans are same. The combinational Trojan delivers the payload once it is triggered, on the other hand, a sequential Trojan needs to be triggered \( R \) times consecutively.

The rest of the paper is organized as follows: an introduction to logic locking along with an overview of different locking techniques and attacks is provided in Section II. The proposed attacks based on hardware Trojan to implement the malicious design modification for the extraction of secret key from any locked circuit is described in Section III. We provide an algorithm for designing an Type-p combinational Trojan considering the set of manufacturing test patterns in Section IV. The number of valid Trojans along with other factors such as area overhead and leakage power for several benchmark circuits are presented in Section V. The future directions are provided in Section VI. Finally, we conclude our paper in Section VII.

II. BACKGROUND

The challenges for protecting a circuit against hardware security threats have been the driving force for the development of different techniques to limit the amount of circuit information that can be recovered by an adversary. Logic locking has emerged as a field of significant interest from the researchers, as it can provide a complete protection against IC overproduction and IP piracy.

The objective of logic locking is to obfuscate the inner details of the circuit and making it infeasible for an adversary to reconstruct the original netlist. Logic Locking hides the functionality of the circuit by inserting additional logic gates into the original design, which we termed as key gates. In addition to the original inputs, the locked circuit needs secret key inputs to key gates from an on-chip tamper-proof memory (see Figure 2(a) for details). The correct functionality of the design is obtained when the key inputs receive the proper secret key value. Applying invalid key to the key gates would result in incorrect functionality of the locked design. Note that for a secure locked circuit, the design details cannot be recovered using reverse engineering.

Different logic locking methods were devised over the years and can be categorized into three different categories. First, XOR-based logic locking, shown in Figure 1(c), have received much attention due to its simplicity. In this technique, a set of XOR or XNOR gates are inserted as key gates [2], [8], [9], [15], [21], [23], [26]. The secret key is stored in tamper-proof memory (TM), and connections are made from TM to the key gates. Second, in MUX-based logic locking technique [27], [28], multiplexers (MUX) are inserted so that one of its input is correct, which is the original net of the circuit. The other input of the MUX is incorrect, which is a dummy net randomly selected from the netlist. This technique is shown in Figure 1(d). The select signal of the MUX is associated with the key bit from the tamper-proof memory. The correct signal goes through the MUX upon applying valid key value, otherwise incorrect signal propagates in the netlist. Third, in LUT-based logic locking, [7], [29], [30], shown in Figure 1(e), a look-up table with several key inputs is used to lock the netlist. The LUTs replace a combinational logic in the design making it difficult to predict the output as it depends on several different key values.

The research community has proposed several attacks to exploit the security vulnerability on a logic locked circuit. Subramanyan et al. [31] first showed that a locked circuit can be broken using Boolean Satisfiability (SAT) analysis. The SAT attack algorithm, attributed as oracle-guided attack, requires a locked netlist, which can be recovered using reverse engineering and functional chip with a valid key stored/programmed in its tamper-proof memory. In this attack, an adversary can query an activated chip and observe the response. Note that SAT attack requires an access to the internal nodes of circuit through the scan chains, which is common in today's netlist for implementing Design-for-Testability (DFT) [32]. The SAT attack works iteratively to eliminate incorrect key values from the key space using distinguishing input patterns (DIPs). A DIP
is defined as an input pattern for which two sets of hypothesis keys produce complementary results. By comparing these with the output of an unlocked chip, one set of hypothesis keys is discarded. The SAT attack works efficiently as it discards multiple hypothesis keys in one iteration.

Thereafter, researchers have focused on improving and developing locking techniques to be resilient against the SAT attack. Subsequent work in this direction involved Anti-SAT [23], [33], SARLock [24], TTLock [25], SFLL [26], design-for-security (DFS) architecture [8], [9], [21]. These proposed techniques use one-point functions. SARLock inverts the output of the circuit for one input pattern corresponding to one incorrect key. This input pattern differ for different incorrect keys. Anti-SAT involves two complementary external logic circuits which are supplied with the same key values and same inputs. The output of these two circuits converge into a AND gate whose output is always 0 for the correct key applied, else, it may be 1 that leads to corrupted internal node value in the original netlist to produce incorrect outputs. Anti-SAT, initially, was proven vulnerable to the signal probability skew (SPS) [24] attack and removal/bypass attack [34]. SARLock was broken by Double DIP attack [35], Approximate SAT attack [36] and removal/bypass attack [34].

Due to limitations of SARLock, Yasin et al. developed an improved version of this design and referred as TTLock [25], where the original design itself is modified to produce corrupted/inverted results upon applying incorrect key. However, TTLock provided protection only for single input pattern so to provide more flexibility in the number of protected input patterns, stripped functionality based logic locking (SFLL) [26] was proposed. The design is no longer same as the original design due to stripped parts of the functionality resulting in erroneous output. A separate restore unit is responsible for removing this error in the output supplying correct key values to it. However, Subramanyan et al. has shown recently that SFLL can be defeated through FALL attack [37]. The attack is build on 3 primary steps, namely, structural analysis, functional analysis and key confirmation. The structural analysis is performed to identify the gates that are output of cube stripping function in SFLL. After identification of these candidate gates, the functional analysis targets the property of cube stripping functions, which results in a set of potential key values. Finally, the key confirmation algorithm identifies the correct key from the set of potential key values.

As the SAT-attack is based on the availability of accessing the internal states of a circuit through the scan chains, Guin et al. proposed placing multiple flip-flops capturing signals controlled by different key bits at the same level of the parallel scan chains, which were used in current test compression methodologies [8]. However, a vulnerability existed in this design, when an adversary performs multi-cycle tests, such as delay tests (transition delay faults and path delay faults) [32]. This leads to the necessity for developing a new design-for-security (DFS) architecture to prevent leaking of the key during any manufacturing tests [9], [21]. This design prevents scanning out the internal states of a design after a chip is being activated and the keys are programmed/stored in the circuit. To the best of our knowledge, this is the only existing solution today, which prevents SAT-based attacks as no attacks have been reported so far.

Apart from SAT based attacks, probing attacks [38] have also shown serious threat to the security of logic locking, where an attacker makes contact with the probes at signal wires in order to extract sensitive information, mainly, the secret key. With the help of focused ion beam (FIB), a powerful circuit editing tool that can mill and deposit material with nanoscale precision, an attacker can circumvent protection mechanisms and reach wires carrying sensitive information. However, the countermeasures reflect the complexity of shield-structure and nanopyramid structures as the defense, making it difficult to perform these attacks [39], [40].

III. PROPOSED TAAL ATTACK FOR EXTRACTING SECRET KEYS

The general hardware security strategy adopted for designing and manufacturing a circuit involves a logic locking, where a chip is unlocked by storing a secret key in the tamper-proof memory. As this secret key is same for all the chips manufactured with the same design, finding this key from one chip undermine the security resulted from logic locking. We show that an adversary can easily extract the key for a chip using our proposed TAAL attacks, which are built on tampering through malicious modification by inserting a hardware Trojan to a locked circuit. In this section, we will present three types of TAAL attacks.

A. Adversarial Model

The adversarial model is given to clearly define the capabilities and intentions of an attacker. In this model, the attacker (adversary) is assumed to be an untrusted foundry and possesses the following:

- The attacker has access to the locked netlist of a circuit. An untrusted foundry has the access to all the layout information, which can be extracted from the GDSII or OASIS file. The netlist can be reconstructed from the layout using reverse engineering with advanced technological tools [41].
- The attacker has the capability to determine the location of the tamper-proof memory. It can also find the location of key gates in a netlist, as it can easily trace the route of the other input of the key gate to the tamper-proof memory.
- The attacker has the capability to tamper a netlist for its malicious intentions through inserting additional circuitry, commonly known as hardware Trojans, about which the SoC designer is unaware.
- The attacker has access to all the manufacturing test (e.g. stuck-at-fault, SAF) patterns. It is common that the production tests are performed at the foundry.

B. T1 Type TAAL Attack

The T1 type TAAL attack is the simplest attack among the other two types, which will be introduced in the successive sections. This attack is beneficial for the attacker who does
C. T2 Type TAAL Attack

Instead of extracting the key directly to the primary output, an adversary can propagate it to the output. In T1 type TAAL Attack, once the Trojan is activated, the raw key values are transferred to primary output, which can raise a suspicion of a design being tampered. T2 type TAAL Attack primarily addresses this shortcomings of T1 type attack, by incorporating logic values in the key, which can easily be separated.

The attack involves tampering a netlist with a Type-3 combinational Trojan. The trigger is constructed using a 3-input AND gate and the payload is delivered to the primary output of the circuit using a 2-input MUX. An adversary can choose a net, whose logic value is impacted by the key gate for the MUX input. In Figure 2(b), net $x_4$ is selected as the MUX input. One can also select $n_3$, however, $n_3$ cannot be selected. The key gate is considered as XOR gate having inputs as secret key ($k$) and $G_1$ gate output. In order to propagate secret key at the output of key gate ($K$), $G_1$ output needs to be specified (either 0 or 1). If $n_1 = 0$, the output of the key gate will be $k$, otherwise it will be $\overline{k}$. Since net $n_4$ is selected for key extraction, input $x_3$ also plays a significant role in key propagation as the complementary value at net $n_2$ can propagate to $n_4$ only when $x_3 = 1$. To launch this attack, an adversary needs to perform the circuit analysis to sensitize the key. This attack requires an adversary to monitor input pattern to extract the correct key. An adversary extracts $x_1$ when $[x_1 \; x_2 \; x_3] = [0 \; 0 \; 1]$ and $k$ otherwise, keeping $x_3$ fixed to logic 1, in the example provided in Figure 2(b). This attack shows the flexibility to identify individual key gate and target secret key through key propagation from consecutive node selection. The dependency of this attack on primary inputs increases the efficiency of T2 type attack where only the adversary has the knowledge about the logical values of these inputs.

D. T3 Type TAAL Attack

A locked netlist typically consists of a large number (e.g., 128) of key gates, the effect of one key may affect the propagation of another key to the primary output. A secure logic locking technique can also insert keys in such a way that an adversary cannot propagate the key information to output using manufacturing tests [15]. In such scenarios, T2 type TAAL attack may be ineffective to extract the key. T3 type TAAL attack is proposed to addresses this limitation encountered for T2 type attack.

Figure 3(a) shows the locked netlist, where the propagation of the key ($k_1$) is prevented by inserting another key ($k_2$). The output of $G_3$ cannot be uniquely determined unless an adversary knows either $k_1$ or $k_2$ and T2 type attack will fail to determine either $k_1$ or $k_2$. It is thus necessary to help propagate one key and then determine the other. Figure 3(b) shows our proposed T3 type TAAL attack, where net $n_5$ is selected to deliver the payload.

Figure 3(b) shows the implantation of T3 type TAAL attack using a Type-3 combinational Trojan. The trigger part for the Trojan can be designed a 3-input AND gate and payload is delivered through 2-input MUX as before. The key ($k_1$) propagation requires setting the node $n_5$ to 1 so that the signal

Figure 2: (a) T1 type TAAL attack, where a Type-3 combinational Trojan is inserted for key extraction directly from the connection between key gate and tamper-proof memory, (b) T2 type TAAL attack, where a Type-3 combinational Trojan is inserted for the secret key extraction.

not intend to gain knowledge regarding the security measures implemented for the circuit. The hardware Trojan assists in extracting out the secret key directly from the tamper-proof memory.

Figure 2(a) shows the proposed modification to launch T1 type TAAL attack. A Type-3 combinational Trojan (see details in Section IV-A) is designed and inserted in the netlist. A 3-input AND gate serves as the trigger and denoted as $T$, whereas, the 2-input multiplexer delivers the payload to the primary output. One input of the multiplexer is the original output of the locked netlist. The other input is connected to the line formed between the key gate ($K$) and the tamper-proof memory as a part of logic locking. Under normal operation for any activated chip, the multiplexer propagates the correct circuit functionality at the output. Once the Trojan gets activated, the output of AND gate becomes 1, which leads to the extraction of secret key through the multiplexer at the output. Note that the required number of multiplexers to extract the complete secret key is dependent on the key size.

The T1 type TAAL attack is very effective as it does not require any knowledge of the circuit netlist. This attack can also be applied to any logic locking techniques without knowing its implementation details as it directly leaks the key to the primary output. As there are no secure measures undertaken in logic locking to protect the connection between the key gates and the tamper-proof memory, any locked circuit can be vulnerable for this attack. Note that an adversary can select any hardware Trojans (combinational or sequential Trojans) of its choice, and one can find the implementation details in Section IV.
A hardware Trojan can be described as intentional modifications in the original netlist of a design for malicious purposes. A Trojan can be inserted into a circuit during its design or manufacturing stages. In this paper, we only consider a Trojan, inserted by an untrusted foundry, which is relevant to logic locking (see adversarial model in Section III-A). As logic locking was proposed to address the treat from an untrusted foundry, it can practically thwart the locking mechanism by obtaining the secret key through inserting a hardware Trojan in the design.

A complete hardware Trojan classification can be found in [13]. In this paper, we only consider combinational and sequential hardware Trojans to demonstrate the attack. A combinational hardware Trojan generally comprises of a trigger and a payload, the detailed modeling can be found in [47]. On the other hand, sequential Trojans have a state element along with the trigger and payload [46]. Any Trojans can be activated through trigger inputs, which can be taken from the primary inputs and/or internal nodes of a circuit, so that manufacturing test patterns cannot trigger a Trojan and remains undetected.

The trigger can be implemented as an AND gate. When a Trojan is activated, the output of this AND gate becomes 1 and it delivers the payload (selection input of the multiplexer shown in Figures 2-3) to the circuit to leak the secret key. The trigger can also be any logic function which provides 1 when activated. Note that a combinational Trojan manifests its effects upon availability of the trigger inputs and effects the original netlist at the payload, on the other hand, sequential Trojan shows its effect after the occurrence of a sequence or a period of time upon triggered.

### IV. Design of Hardware Trojans for TAAL Attacks

A hardware Trojan can be described as intentional modifications in the original netlist of a design for malicious purposes. A Trojan can be inserted into a circuit during its design or manufacturing stages. In this paper, we only consider a Trojan, inserted by an untrusted foundry, which is relevant to logic locking (see adversarial model in Section III-A). As logic locking was proposed to address the threat from an untrusted foundry, it can practically thwart the locking mechanism by obtaining the secret key through inserting a hardware Trojan in the design.

A complete hardware Trojan classification can be found in [13]. In this paper, we only consider combinational and sequential hardware Trojans to demonstrate the attack. A combinational hardware Trojan generally comprises of a trigger and a payload, the detailed modeling can be found in [47]. On the other hand, sequential Trojans have a state element along with the trigger and payload [46]. Any Trojans can be activated through trigger inputs, which can be taken from the primary inputs and/or internal nodes of a circuit, so that manufacturing test patterns cannot trigger a Trojan and remains undetected.

The trigger can be implemented as an AND gate. When a Trojan is activated, the output of this AND gate becomes 1 and it delivers the payload (selection input of the multiplexer shown in Figures 2-3) to the circuit to leak the secret key. The trigger can also be any logic function which provides 1 when activated. Note that a combinational Trojan manifests its effects upon availability of the trigger inputs and affects the original netlist at the payload, on the other hand, sequential Trojan shows its effect after the occurrence of a sequence or a period of time upon triggered.

#### IV. Design of Hardware Trojans for TAAL Attacks

A hardware Trojan can be described as intentional modifications in the original netlist of a design for malicious purposes. A Trojan can be inserted into a circuit during its design or manufacturing stages. In this paper, we only consider a Trojan, inserted by an untrusted foundry, which is relevant to logic locking (see adversarial model in Section III-A). As logic locking was proposed to address the threat from an untrusted foundry, it can practically thwart the locking mechanism by obtaining the secret key through inserting a hardware Trojan in the design.

A complete hardware Trojan classification can be found in [13]. In this paper, we only consider combinational and sequential hardware Trojans to demonstrate the attack. A combinational hardware Trojan generally comprises of a trigger and a payload, the detailed modeling can be found in [47]. On the other hand, sequential Trojans have a state element along with the trigger and payload [46]. Any Trojans can be activated through trigger inputs, which can be taken from the primary inputs and/or internal nodes of a circuit, so that manufacturing test patterns cannot trigger a Trojan and remains undetected.

The trigger can be implemented as an AND gate. When a Trojan is activated, the output of this AND gate becomes 1 and it delivers the payload (selection input of the multiplexer shown in Figures 2-3) to the circuit to leak the secret key. The trigger can also be any logic function which provides 1 when activated. Note that a combinational Trojan manifests its effects upon availability of the trigger inputs and affects the original netlist at the payload, on the other hand, sequential Trojan shows its effect after the occurrence of a sequence or a period of time upon triggered.

#### IV. Design of Hardware Trojans for TAAL Attacks

A hardware Trojan can be described as intentional modifications in the original netlist of a design for malicious purposes. A Trojan can be inserted into a circuit during its design or manufacturing stages. In this paper, we only consider a Trojan, inserted by an untrusted foundry, which is relevant to logic locking (see adversarial model in Section III-A). As logic locking was proposed to address the threat from an untrusted foundry, it can practically thwart the locking mechanism by obtaining the secret key through inserting a hardware Trojan in the design.

A complete hardware Trojan classification can be found in [13]. In this paper, we only consider combinational and sequential hardware Trojans to demonstrate the attack. A combinational hardware Trojan generally comprises of a trigger and a payload, the detailed modeling can be found in [47]. On the other hand, sequential Trojans have a state element along with the trigger and payload [46]. Any Trojans can be activated through trigger inputs, which can be taken from the primary inputs and/or internal nodes of a circuit, so that manufacturing test patterns cannot trigger a Trojan and remains undetected.

The trigger can be implemented as an AND gate. When a Trojan is activated, the output of this AND gate becomes 1 and it delivers the payload (selection input of the multiplexer shown in Figures 2-3) to the circuit to leak the secret key. The trigger can also be any logic function which provides 1 when activated. Note that a combinational Trojan manifests its effects upon availability of the trigger inputs and affects the original netlist at the payload, on the other hand, sequential Trojan shows its effect after the occurrence of a sequence or a period of time upon triggered.
inputs and/or internal nodes, which are not affected by the key gates. If such a node selected as a trigger input, an adversary cannot activate a Trojan as it does not know this secret key, and thus the internal signal value for an activation pattern. We call this pattern as hardware Trojan activation pattern (HTAP).

The payload of the Trojan (a MUX) can be delivered to a location described in Figures 4 for launching our proposed TAAL attack.

Let us determine the number of Trojans, one can insert in a design to extract the secret key. This is basically a selection problem, where an adversary selects $p$ nodes as the trigger inputs from $N$ nodes of a circuit so that the Trojan is not activated during the manufacturing/production tests. The value of $N$ can be determined based on the following equation:

$$ N = PI + G + F - M $$

where,

- **PI**: Number of primary inputs
- **G**: Number of gates
- **F**: Number of fanout branches
- **M**: Number of lines impacted by the key gates

An upper bound of all possible Type-$p$ Trojans ($AT_p$) can be given by:

$$ AT_p = \binom{N}{p} \times 2^p $$

The right hand side of Equation 2 constitutes of two products. The first one represents all possible combinations to select $p$ lines from $N$. The second one denotes the trigger combinations, as one line can be applied directly or inverted to the trigger input. Note that the actual number of Trojans (denoted as $VT_p$) can be less than $AT_p$ as few of them can be detected by the manufacturing test patterns (e.g., stuck-at fault patterns), and few may not be triggered from the primary inputs. However, for a reasonable size circuit, $AT_p$ and $VT_p$ are comparable.

Figure 4 shows an example of TAAL: $T1$ type using a Type-4 Trojan inserted in the netlist. The circuit has five primary inputs ($PI$). The SoC designer can generate test patterns considering the key as input (the pattern generation is described in detail in [9], [21]). To detect all the stuck-at faults (SAFs), seven test patterns (e.g., $P = \{P_1, P_2, \ldots, P_7\}$) are required and they are generated using Synopsys TetraMax [48] ATPG tool. To avoid a Trojan being activated by these manufacturing test patterns, the trigger of the Trojan must remain quiet for all these input patterns. A hardware Trojan activation pattern is selected, where $HTAP = (X \ 0 \ 0 \ 1 \ 1 \ 0)^T \notin P$. As the logic values of nodes $n_2, n_4$, and $n_5$ are impacted by the key, $k$, these nodes are excluded in designing the Trojan. If one of these nodes are selected, an adversary may not activate the trigger as it does not know the key value. The upper bound of all possible Type-4 Trojans ($AT_4$) can be given by

$$ AT_4 = \binom{7}{4} \times 2^4 = 560 $$

where, $N = 5 + 5 - 3 = 7$.

Out of these 560 possible Trojans 188 will be detected by the test patterns $P$. The remaining 372 will be treated as valid Trojans, and an adversary can select one of them. In Figure 4, $x_1, x_2, x_3$ and $x_5$ are selected as the trigger. Similarly, one can also design other types (Type-1 through Type-6) of Trojans to launch TAAL attacks. Note that the Trojan needs to be quiet during normal operations, so that no functional errors are observed at the primary output. An adversary can select rare nodes, whose value do not become identical with trigger pattern very often under continuous/normal operation of the IC, for the trigger inputs while designing a Trojan. One can perform controllability and observability analysis [32] to find such rare nodes, and then select them as the trigger inputs. However, we do not include such analysis, as including of rare nodes in the trigger for a sequential hardware Trojan is not a hard requirement. In this paper, a sequential Trojan is modelled using a combinational Trojan that needs to be triggered $R$ times consecutively.

To design a combinational hardware Trojan, an automated process is developed. Algorithm [1] provides steps to be followed for designing a Type-$p$ Trojan that eludes activation during the manufacturing test. The inputs of the algorithm are locked netlist ($C$), $M$ production/manufacturing test patterns ($P = \{P_1, P_2, \ldots, P_M\}$) and the Trojan type ($p$). The algorithm results the hardware Trojan activation pattern (HTAP) and the trigger inputs. Initially, it reads the Trojan-free locked netlist ($C$) and the set of manufacturing test patterns ($P$) (Lines 1-2). Logic simulation is performed using these patterns, and
Algorithm 1: Design of Type-p Trojan

input:  Locked Netlist (C), test pattern set (P), Type-p Trojan
output: Hardware Trojan activation pattern (HTAP),
        Trigger inputs (T)
1 Read the locked netlist (C);
2 Read production test patterns (P);
3 Perform logic simulation using P to form a matrix (A) of all the internal node values;
4 Select a hardware Trojan activation pattern (HTAP), where HTAP \notin P;
5 Perform logic simulation using HTAP and form a
   matrix (H) of all the internal node values;
6 Select \( p \) random nodes that are not affected by the key gates of C for the trigger inputs;
7 Construct a new matrix \( A_p \) that corresponds to the
   trigger locations for all test patterns;
8 Construct a new vector \( H_p \) that corresponds to the trigger locations for HTAP ;
9 if \( H_p \notin A_p \) then
10 \hspace{1em} Choose selected \( p \) nodes as trigger, T;
11 else
12 \hspace{1em} Discard selected \( p \) nodes, as it would activate the
   Trojan during tests;
13 \hspace{1em} Go to Step 6;
14 end
15 Report HTAP and T;

store the internal node values in a matrix, A (Line 3). It is not necessary to store the nodes those are impacted by the key gates, as their values will be unknown (Xs) during the simulation. Note that A is a \( N \times M \) matrix. A hardware Trojan activation pattern of an adversary’s choice is selected (Line 4). Similarly, matrix H is formed as the result of logic simulation using HTAP (Line 5). Here, H is a \( N \times 1 \) vector. To select the trigger inputs, one can select \( p \) random nodes that are not affected by the key gates of the locked circuit (Line 6). To perform the search whether the trigger values are presented in the production set, matrix \( A_p \) and vector \( H_p \) are constructed (Lines 7-8). If \( H_p \) is not in \( A_p \), the trigger (T) is selected (Line 10), otherwise drop the selected \( p \) locations as the Trojan will be activated during the production tests (Lines 12-13) and new \( p \) locations are selected (Line 6). Finally, the algorithm reports HTAP and T (Line 15).

B. Design for a Sequential Hardware Trojan

A sequential Trojan modifies the functionality of a circuit until a specified time has elapsed after the trigger condition is satisfied. However, in this paper we designed a sequential Trojan that needs to be triggered \( R \) times to deliver the payload. We designed a sequential Trojan in this way, so that it can be modelled using a combinational Trojan, described in detail in the previous section.

A sequential Trojan also consists of a trigger and payload similar to a combinational Trojan. Additionally, the trigger part contains state elements that ascertain the payload in future time. In our sequential Trojan design, a \( R \)-bit counter is implemented as the state elements. This counter is enabled (en) once the trigger condition is fulfilled, i.e., the output of the \( p \)-input AND gate becomes 1. The counter increments by one unit, every-time the Trojan is triggered using the Trojan activation pattern (HTAP). The Trojan delivers at the payload (MUX) only after reaching the maximum counter value (\( R \)). The circuit tampered with Sequential Trojan will show the intended malfunction, key extraction in TAAL attacks, only upon applying activation pattern successively \( R \)-times to the circuit.

Figure 5(a) shows the TAAL attack using a sequential Trojan. The trigger consists of a \( p \)-input AND gate and a \( R \)-bit counter. The finite-state machine (FSM) of the counter is shown in Figure 5(b). The FSM goes to next states, when \( en = 1 \), otherwise, it returns to the initial state, \( S_0 \). The counter produces an output of 1, once \( en \) is hold to 1 consecutively \( R \) clock cycles, as it takes \( (R - 1) \) cycles to reach \( S_{R-1} \). Note that this sequential Trojan can be modelled as \( R \) number of combinational Trojans. An adversary can also design a different sequential Trojan, already in the literature, to launch the TAAL attack. The sequential Trojan increases the complexity compared to a combinational Trojan as it manifests its effect to the payload only after sequence of repeated application of trigger inputs. Only the adversary has the knowledge regarding the maximum counter value making it very difficult for detection.
V. Analysis

The hardware Trojans presented in Section IV pose a unique challenge to the SoC designers for securing their designs. In this section, we show that an adversary can implement a Trojan in a very large number of ways, and it is practically infeasible to detect all of them with absolute certainty. We choose six benchmark circuits from ISCAS’85 benchmark suites [49] to show the complexity of Trojan detection even for these small benchmark circuits.

Table I shows the design details for different locked benchmark circuits. The number of logic gates and key size for these circuits are shown in Columns 2 and 3, respectively. The number of key bits are selected in such a way that the total area overhead does not exceed 5%. However, for an industrial design with millions of gates, the key gates will merely add any overhead. Column 4 represents the total number of nets in these circuits (see Equation 1). The number of nets that are not affected by key gates is shown in Column 5. Note that these nets cannot be selected for trigger inputs. The manufacturing test patterns are generated using Synopsys TetraMax Automatic Test Pattern Generation (ATPG) tool [48] with targeted 100% fault coverage (Columns 6-7). For C432 benchmark, we insert 30 key gates randomly in the netlist with 160 logic gates. There are 349 nets in the netlist, out of which 233 nets can be selected for Trojan trigger as the remaining nets are affected by the key gates. The TetraMax ATPG tool generates 58 stack-at fault patterns and reports 100% fault coverage. These test patterns will be used by an adversary to design the Trojans such that they are not activated during the manufacturing tests. Similar analysis can be performed for all other benchmark circuits through the details mentioned in respective rows.

Table II shows the number of combinational hardware Trojans that can be designed to perform TAAL attacks (mentioned in Section III) for different benchmark circuits. The upper bound (see Equation 2) for all possible Trojans that can be inserted in the circuit is denoted in Columns 2, 4 and 6. Out of all possible Trojans, the valid Trojans that will not be detected during manufacturing tests, are shown in Columns 3, 5 and 7. For C432 benchmark circuit, the total number of Type-2 Trojans is $1.08 \times 10^5$, whereas, the number of valid Trojans is $1.0 \times 10^5$. The number of Trojans increases exponentially with the increase of the Trojan type ($p$). Note that $AT_p$ and $VT_p$ are in the same order, which gives an adversary to select a Trojan of its choice from a large collection. It is worthwhile to mention that an adversary needs to choose a Trojan whose triggers are selected from the rare nodes such that it does not get activated during normal operation. However, it is not necessary to impose this condition for designing a sequential Trojan, as it is highly unlikely that a particular trigger condition will arrive $R$ times consecutively during the normal operation of a chip.

VI. Future Research Direction for Secure Logic Locking

The security of a logic locking technique can be tied together with the hardware Trojan detection problem. Developing a SAF-registrant logic locking is not sufficient enough to prevent IC overproduction or to protect IPs. It is required to address the detection of Trojans inserted at an untrusted manufacturing site. Researchers have already proposed different techniques to detect and prevent hardware Trojans. The detection methods can be grouped into two different categories, such as, logic testing [50]–[54], and side-channel analysis [55]–[60]. On the other hand, prevention methods can be categorized as design-for-trust measures [61]–[65] and split manufacturing [66]–[68].

Logic testing by applying stimuli to primary inputs (PIs) and observe responses at primary outputs (POs) can be used to detect these Trojans [46], [50], [52]–[54], [69]. Decision is being made whether a chip is tampered with a hardware Trojan by observing a mismatch between the observed and expected responses. Note that the accuracy of the detection process does not depend on the manufacturing process variations. However, the detection will be extremely difficult as it is practically impossible to detect all types of combinational Trojans. In addition, it is not feasible to trigger a sequential Trojan, as it requires to apply the same trigger pattern at the input $R$ times.

Side channel information, such as, power [70], temperature [71], delay [72], and radiation [73] can be used to detect a hardware Trojan. These detection methods rely on the availability of Trojan-free golden circuits for creating Trojan free signature. It can be very difficult to acquire a golden sample as all the chips may have Trojans. In addition, process and environmental variations may mask the side channel leakage, when a Trojan circuitry is small.

While dedicated towards hardware Trojan detection, researchers propose different measures to prevent a Trojan being inserted into the design in the first place. These solutions involved characterization of ring-oscillator [62], shadow registers [59], and delay elements [74] to detect the delay deviation caused by hardware Trojans. Reducing the rare signal in the circuitry is another proposed method for designers to reduce the risk of being implanted with a Trojan [63], [75]. Camouflage fill techniques [76], [77] to create indistinguishable layouts for different gates by adding dummy contacts and connections can prevent the attacker from extracting a correct gate-level netlist of a circuit for Trojan insertion. Xiao et al. proposed to fill all the unused spaces using filler cells so that an untrusted foundry cannot insert a Trojan [64], [78]. However, this direction still lacks any firm solution as more emphasis is observed in Trojan detection.

Split Manufacturing can be an effective way to thwart Hardware Trojan insertion at an untrusted foundry. In split manufacturing, the production of ICs is carried out in two different foundries [79]. The design is divided into two parts – Front End of Line (FEOL) and Back End of Line (BEOL). An untrusted foundry is provided with the FEOL design, which contains partial information regarding the design that requires complex steps for fabricating and involves higher cost. Fabrication of BEOL does not incorporate complex fabrication steps and can be done by a smaller trusted foundry. The untrusted foundry sends the fabricated wafers directly to the smaller foundry for the complete fabrication. This way the untrusted foundry can be restricted to make any Trojan based modification as it does not have the complete information...
Table I: Circuit parameters.

| Benchmarks | # Gates | Key Size (|K|) | # Total Lines (N + M) | # Net Lines (N) | # Test Patterns | Fault coverage |
|------------|---------|---------|----------------------|----------------|----------------|---------------|
| C432       | 160     | 30      | 349                  | 233            | 58             | 100%          |
| C499       | 202     | 30      | 491                  | 226            | 78             | 100%          |
| C880       | 383     | 30      | 594                  | 350            | 86             | 100%          |
| C1908      | 880     | 30      | 552                  | 223            | 83             | 100%          |
| C3540      | 1669    | 83      | 1826                 | 1114           | 173            | 100%          |
| C6288      | 2416    | 128     | 5621                 | 1335           | 77             | 100%          |

Table II: Number of hardware Trojans for launching TAAL attacks.

| Benchmarks | Type-2 Trojan | Type-3 Trojan | Type-4 Trojan |
|------------|---------------|---------------|---------------|
|            | AT_2          | VT_2          | AT_3          | VT_3          | AT_4          | VT_4          |
| C432       | 1.08 × 10^9  | 1.04 × 10^10  | 1.66 × 10^1  | 1.43 × 10^9  | 1.91 × 10^9  | 1.34 × 10^9  |
| C499       | 1.02 × 10^2  | 0.27 × 10^3    | 1.52 × 10^2  | 2.11 × 10^6  | 1.69 × 10^9  | 1.21 × 10^9  |
| C880       | 2.44 × 10^4  | 2.25 × 10^5    | 5.67 × 10^4  | 4.80 × 10^10 | 9.83 × 10^9  | 7.35 × 10^9  |
| C1908      | 0.99 × 10^7  | 0.96 × 10^8    | 1.46 × 10^7  | 1.33 × 10^7  | 1.60 × 10^9  | 1.27 × 10^9  |
| C3540      | 2.48 × 10^9  | 2.35 × 10^10   | 1.84 × 10^9  | 1.57 × 10^9  | 1.02 × 10^12 | 0.74 × 10^12 |
| C6288      | 3.56 × 10^7  | 3.50 × 10^8    | 3.17 × 10^9  | 3.00 × 10^9  | 2.11 × 10^12 | 1.82 × 10^12 |

Regarding the design. However, several attacks undermining the security achieved through split manufacturing have also been proposed in past [80], [82].

Recent research contributions showed that machine learning and image processing can also be incorporated to detect hardware Trojans in the chip. Vashistha et al. presented Trojan scanner [83], which uses a trusted GDSII layout (golden layout) and scanning electron microscope (SEM) images to identify the malicious modifications made in the netlist during the manufacturing of a circuit. A unique descriptor for each type of gate is prepared based on different features using computer vision algorithms along with machine-learning model of a golden layout and SEM images of an IC under authentication. These descriptors, when compared to each other can detect any modifications either in the form of additional gates or modified gates which might raise the suspicion for a potential hardware Trojan. Moreover, Trojan scanner also presents the trade-off between the accuracy and SEM parameters. The authors demonstrated the effectiveness of the scheme using a smart card die as test sample (generally manufactured with 90 nm technology). It is yet to be validated its effectiveness of detection when a chip is fabricated using recent technology nodes (10 nm and beyond).

Despite significant research have been performed on detecting hardware Trojans, we still lack efficient and accurate methods for modeling them and generating tests for their detection. Once the detection of hardware Trojans is ensured, an SoC designer can choose a SAT-resistant logic locking to prevent IC overproduction and IP piracy.

VII. CONCLUSION

In this paper, we have demonstrated the vulnerability of logic locking techniques through a set of tampering attacks with hardware Trojans. Three types of proposed TAAL attacks can defeat any logic locking techniques that rely on storing the secret key in a tamper-proof memory. In T1 type TAAL Attack, we showed how an adversary can extract the key from a locked netlist without knowing the details of logic locking technique used to protect the circuit. For T2 type and T3 type TAAL attacks, the complexity of detecting an attack has been improved. Only the attacker has the knowledge about the specific values that can lead to key extraction, increasing the identification of a TAAL attack. To launch a TAAL attack, we develop models for combinational and sequential hardware Trojans. We also proposed an algorithm to design a hardware Trojan that cannot be detected by manufacturing tests. The results depict the range of Trojans that can be selected by an adversary, which has a very high order of magnitude. Finally, we describe relevant detection and avoidance strategies for hardware Trojans to make logic locking secure.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under grant number CNS-1755733.

REFERENCES

[1] Age Yeh, “Trends in the global IC design service market,” DIGITIMES Research, 2012.
[2] J. A. Roy, F. Koushanfar, and I. L. Markov, “EPIC: Ending piracy of integrated circuits,” in Proceedings of the conference on Design, automation and test in Europe, 2008, pp. 1069–1074.
[3] Y. Alkabani and F. Koushanfar, “Active Hardware Metering for Intellectual Property Protection and Security,” in USENIX security symposium, 2007, pp. 291–306.
[4] R. S. Chakrabarty and S. Bhania, “Hardware protection and authentication through netlist level obfuscation,” in Proc. of IEEE/ACM International Conference on Computer-Aided Design, 2008, pp. 674–677.
[5] Y. Alkabani, F. Koushanfar, and M. Potkonjak, “Remote activation of ICs for piracy prevention and digital right management,” in Proc. of IEEE/ACM Int. conf. on Computer-aided design, 2007, pp. 674–677.
[6] I. Huang and I. Lach, “IC activation and user authentication for security-sensitive systems,” in IEEE International Workshop on Hardware-Oriented Security and Trust, 2008, pp. 76–80.
[7] A. Baumgartner, A. Tyagi, and J. Zambreno, “Preventing IC piracy using reconfigurable logic barriers,” IEEE Design & Test of Computers, vol. 27, no. 1, pp. 66–75, 2010.
[8] U. Guin, Q. Shi, D. Forte, and M. M. Tehranipoor, “FORTIS: a comprehensive solution for establishing forward trust for protecting IPs and ICs,” ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 21, no. 4, p. 63, 2016.
