TeSR: A Robust Temporal Self-Referencing Approach for Hardware Trojan Detection

Seetharam Narasimhan[1], Xinmu Wang[1], Dongdong Du[1], Rajat Subhra Chakraborty[2], and Swarup Bhunia[1]
[1]Case Western Reserve University, Cleveland, Ohio, USA
[2]Indian Institute of Technology, Kharagpur, West Bengal, India
Email: *sxn124@case.edu*

***Abstract*—Malicious modification of integrated circuits, referred to as *Hardware Trojans*, in untrusted fabrication facility has emerged as a major security threat. Logic testing approaches are not very effective for detecting large sequential Trojans which require multiple state transitions often triggered by rare circuit events in order to activate and cause malfunction. On the other hand, side-channel analysis has emerged as an effective approach for detection of such large sequential Trojans. However, existing side-channel approaches suffer from large reduction in detection sensitivity with increasing process variations or decreasing Trojan size. In this paper, we propose TeSR, a Temporal Self-Referencing approach that compares the current signature of a chip at two different time windows to completely eliminate the effect of process noise, thus providing high detection sensitivity for Trojans of varying size. Furthermore, unlike existing approaches, it does not require golden chip instances as a reference. Simulation results for three complex designs and three representative sequential Trojan circuits demonstrate the effectiveness of the approach under large inter- and intra-die process variations.**

***Index Terms*—Hardware Trojan, side-channel analysis, self-referencing, Trust in IC.**

I. Introduction

An emerging security concern with integrated circuit (IC) involves its malicious modification during fabrication [1] in untrusted foundry. Such malicious hardware modifications, also referred to as *Hardware Trojans*, can give rise to undesired functional behavior of a chip, or provide covert channels or *back doors* through which sensitive information can be leaked. Conventional structural and functional testing fails to reliably detect these Trojans due to their stealthy nature and inordinately large number of instances an adversary can exploit. Hardware Trojan circuits can be either *combinational* or *sequential* [2] in nature. A *combinational Trojan* depends on the occurrence of rare logic values at one or more internal circuit nodes to trigger, while a *sequential Trojan* acts as a *time-bomb*, exhibiting its malicious effect due to a sequence of rare events after long period of operation. Fig. 1(a) shows a generic model for sequential Trojan. Examples of sequential Trojan circuits are k-bit synchronous counter, as shown in Fig. 1(b) and specially-crafted Finite State Machine (FSM) which is triggered by rare events in the internal circuit nodes, as shown in Fig. 1(c). Trojan activation condition is referred as *Trigger condition*, while the node that can be affected when the Trojan is triggered is referred as *payload*. The state transitions are caused by *partial trigger conditions* (PTC). A sequential Trojan with a *passive* payload [3], consists of a Linear Feedback Shift Register (LFSR) which is used to leak the secret key from cryptographic hardware by aiding side-channel attacks, as shown in Fig. 1(d).

Sequential Trojans can be extremely hard to detect using logic testing approaches [2] due to the difficulty of satisfying the rare sequence of state transitions in a Trojan that leads to modification of its payload. Logic testing approaches are more effective for detecting combinational or small sequential Trojans. On the other hand, side-channel analysis is based on noting the Trojan effect on physical side-channel parameters such as current [4] or delay [5]. These approaches do not require Trojan activation or the propagation of its malicious effect to the primary outputs. However, they suffer from reduced sensitivity with increasing inter-die and intra-die *process variation* effects [6], which can mask the Trojan effect leading to false positive/negative decisions. In [7], the authors use current measurement from multiple ports along with calibration techniques and statistical analysis to alleviate process variations. Correlation between multiple side-channel parameters [8] or the same parameter measured from different regions of the chip [9] can be used to calibrate inter-die process noise. Other methods propose region-based test vector generation [10] to increase sensitivity and gate-level charac-

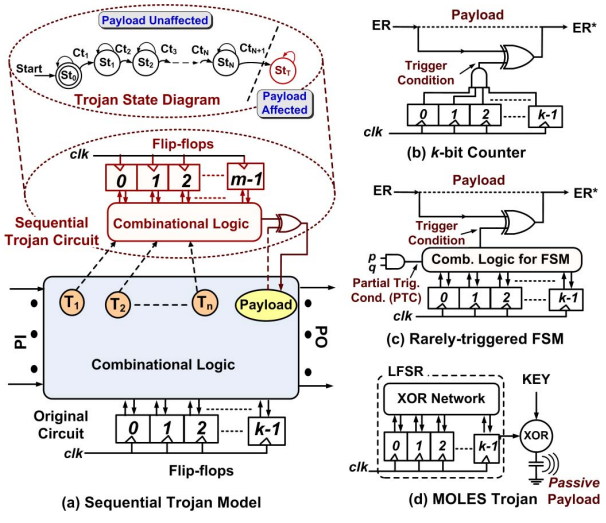
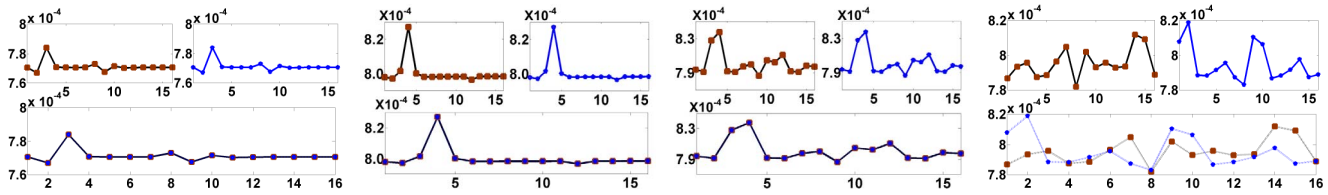


Fig. 1. (a) Sequential Trojan model and examples: (b) Synchronous Counter, (c) Rarely-triggered Finite State Machine (FSM), (d) *MOLES* Trojan [3].



(a) Non-infected, Nominal (b) Non-infected, Inter-die (c) Non-infected, Intra-die (d) Infected, Nominal

Fig. 2. Effectiveness of temporal self-referencing in detecting sequential Trojans in presence of process variations.

terization [11] of leakage and delay parameters for Trojan detection. Existing side-channel methods cannot completely mitigate the influence of process variations - particularly, the within-die variations. Moreover, they rely on the availability of golden ICs (which can be obtained by destructive testing of a sample of test ICs) or complete characterization of the golden design, which can be of prohibitive complexity.

In this paper, we present a novel side-channel analysis approach referred as **Te**mporal **S**elf-**R**eferencing or **TeSR** for sequential Trojan detection that can eliminate the effects of process variation, both inter-die and intra-die (both systematic and random components). It also avoids the requirement of a reference or golden IC by comparing a chip's transient current signature with itself - but at a different time window. TeSR focuses on identifying the sequential Trojans, which typically represent a greater threat than their combinational counterparts, since an intelligent attacker can create a complex Trojan with extremely rare trigger conditions using just few state elements (e.g. flip-flops). The main insight is that when a Trojan-free circuit is made to undergo the same set of state transitions multiple times, the transient current "signature" should remain constant over different time windows. However, in a Trojan-infected circuit, the current signature varies over multiple time windows for the same set of state transitions of the original circuit, due to uncorrelated state transitions in the Trojan. This paper also provides a test generation approach for maximizing switching activity in arbitrary Trojan circuits, triggered by rare node conditions. Effectiveness of the proposed approach is verified with several large IP cores for three representative sequential Trojan circuits.

II. Motivation for Temporal Self-Referencing

As a motivational example of TeSR-based Trojan detection, we simulated a 32-bit DLX processor (with $\sim 20,000$ logic gates) in HSPICE using 70nm Predictive Technology Model (PTM) [12]. Test vector sets are designed to fill the pipeline with repeated "NOP" or "ADD" instructions, causing controlled activity in one pipeline stage at-a-time. Multiple instances of the processor were considered - non-infected and infected, at different process corners, to demonstrate the existence of time-invariant (but process-dependent) signature in each non-infected IC. The Trojan was modeled as a free-running synchronous 8-bit binary counter (Fig. 1(b)), which causes malfunction in a payload node upon reaching the maximum count. The measured side-channel parameter is the average transient supply current in each clock cycle. We used *Monte Carlo* simulations in *HSPICE* with $\pm 20\%$ variations in inter-die transistor *threshold voltage* V_T and intra-die variations having a standard deviation (σ) of 10%.

Fig. 2(a) shows the cycle-by-cycle average transient current trace of the DLX circuit for two windows, where it was repeatedly brought to the same state and made to go through the same set of state transitions. This current trace can be clearly distinguished by its repetitive nature. It forms a "current signature" for this state transition sequence, as seen in the bottom plot of Fig. 2(a), where the current signatures from the two windows are superimposed on each other. In Fig. 2(b), the current signatures for the same two windows are plotted for a non-infected die at a different inter-die process corner. Process variations cause considerable change in the golden signature from chip-to-chip, but the signature for the same IC instance remains time-invariant. This holds true even under intra-die variations, as seen in Fig. 2(c). Now, let us consider the current signatures for a Trojan-infected DLX circuit in Fig. 2(d). Note that the average current value of the Trojan chip at the nominal corner is similar to that of a non-infected chip at a different process corner, thus its effect is masked by process noise. However, since the Trojan state machine undergoes a set of state transitions uncorrelated to the original circuit, the current traces in the two time windows differ substantially. This example motivates the use of temporal self-referencing as a high-sensitivity Trojan detection scheme.

III. Methodology

The major steps of the temporal self-referencing methodology are shown in Fig. 3. It involves both *test generation* and current measurement-based *circuit characterization*.

Test Generation: Our test generation methodology integrates both functional testing and side-channel analysis aspects in order to satisfy the partial trigger conditions caused by rare node values while causing maximum switching activity within the Trojan. The main steps for test pattern generation are shown in Fig. 3. The given circuit is first decomposed into non-overlapping functional partitions or modules $\{M_i\}$ to decrease the complexity of test vector generation while increasing detection sensitivity of the side-channel parameter [9]. For each module M_i, we employ a statistical test pattern generation approach. First, a list of rarely triggered internal nodes in the circuit netlist is identified through simulations with large number of random patterns, along with their respective rare values (logic-0 or logic-1). Then, a compact testset is generated that triggers each rare node to its rare value at

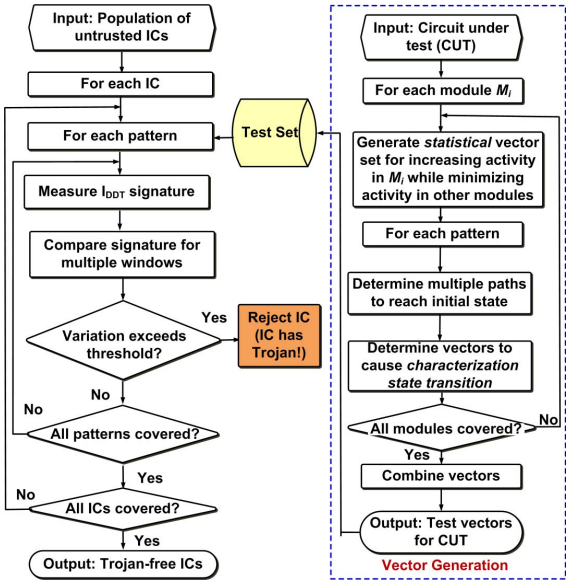


Fig. 3. The major steps of the TeSR approach for Trojan detection.

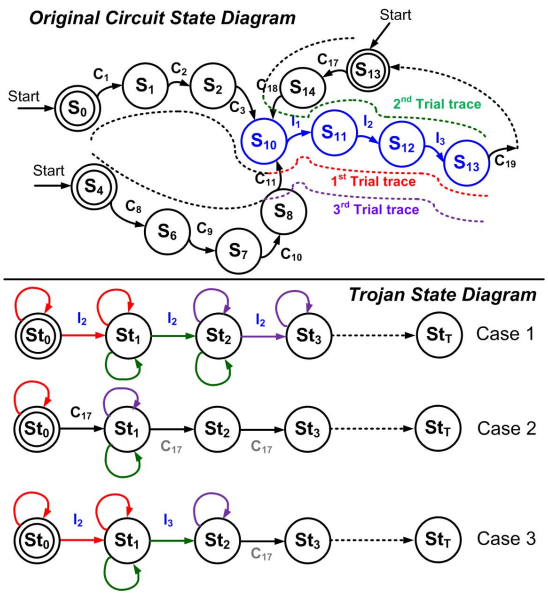


Fig. 4. State transitions of the original design and inserted sequential Trojan for different trials of test vector application for different Trojan cases.

least N times, a technique that has superior trigger coverage compared to random test patterns [2].

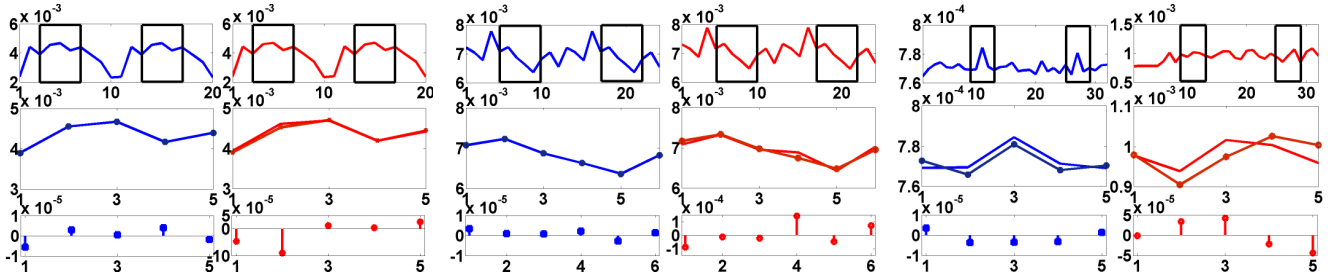
From this compact set, a subset $\{I_j^1\}$ is determined that ensures that the circuit is brought to the same pre-determined state S_j (e.g. S_{10} in Fig. 4) to initiate the current signature characterization for a given region. Existing "Design for Testability" (DfT) techniques such as"Full-scan" or "partial scan" can be used to initialize the circuit to a state S_{jk} which is close to S_j. Once the circuit is in the desired starting state S_j, the set of test vectors $\{I_j^2\}$ is applied which takes the circuit through a fixed set of state transitions (e.g. S_{11} through S_{13} in Fig. 4) in order to produce the characteristic current trace. Hence, overall a pattern set corresponding to a pre-defined state S_j and region $\{M_i\}$ is $V_j^i = \bigcup_k (S_{jk}, \{I_{jk}^1\}, \{I_j^2\})$. Multiple such S_j states are considered, and for each state S_j multiple paths to reach the state S_j are determined by *sequential justification*.

Circuit Characterization: From the generated test set, the sequence of test vectors $\{I_j^1\}$ are applied which takes the circuit to the state S_j, followed by the set $\{I_j^2\}$ that makes the circuit go through a fixed set of state transitions in order to produce the characteristic current trace. It is desirable to have the lengths of the different paths leading to state S_j to be different (ideally, mutually prime). Otherwise, a free-running Trojan state machine might be synchronized to the *test control* (TC) signal, or to the *reset* signal in a non-scan design. An arbitrary FSM Trojan may either undergo one or more state transitions or stay in the initial state, depending on how many of the vectors caused the Trojan PTCs. Let us consider three consecutive trials which are shown by red, green, and purple traces, in the state diagrams for three different FSM Trojans in Fig. 4. If any of the vectors $\{I_j^2\}$ during the first trial and the vectors $\{I_j^1\}$ for re-initializing original circuit state cause at least one state transition of the Trojan state machine, it will be in different states during two consecutive trials. The combinational switching activity with or without accompanying state transitions in the Trojan, always depends on its current state. In order to cover different possible Trojan activation conditions within a region, the signatures have to be compared over multiple consecutive trials.

The current signature is computed by taking the average of the transient current waveform for each cycle. The *difference metric* for comparing the current signature of two windows is taken as the point-wise Euclidean distance between the two current signatures. If one or more of the current traces differ from the average current trace over multiple windows by a pre-defined *noise threshold*, the IC is inferred to contain a Trojan. This noise threshold value can be obtained by taking multiple current measurements with constant activity (reset state) to characterize the background noise in the measurement setup. Unlike other side-channel Trojan detection approaches, we do not require one or more golden ICs to determine the threshold or to calibrate process or measurement noise.

IV. Results

We used three test circuits to validate the proposed Trojan detection approach: 1) an AES cipher circuit, 2) a 32-bit pipelined Integer Execution Unit (IEU) and 3) a 32-bit DLX processor. We introduced three types of sequential Trojan circuits (see Fig. 1). Fig. 5(a) shows the plot of average current over each clock cycle for the IEU circuit with (red) and without (blue) an 8-bit counter Trojan, with the signature highlighted using black rectangles. The superimposed current signatures and their difference are also plotted. It can be clearly observed that there is a significant difference in the signatures for the two windows due to presence of Trojan. Similar waveforms are plotted for different test circuit-Trojan combinations in Fig. 5(b) and Fig. 5(c), respectively. Note that we validate each



(a) IEU with an 8-bit counter (b) AES with a MOLES Trojan (c) DLX with a FSM Trojan

Fig. 5. Temporal self-referencing can be used to detect different types of sequential Trojans in different designs. Note that the difference metric (bottom subplot) for non-infected chips (in blue) is much less than for Trojan-infected chips (in red).

TABLE I
DIFFERENCE METRIC AND TEST LENGTH COMPARISONS.

	Test Length	Difference Metric (μA)			
		No Trojan	Counter	FSM	LFSR
IEU	752	2.76	47.26	214.30	89.88
AES	1161	3.11	87.09	215.30	78.28
DLX	605	2.96	4.10	33.90	33.63

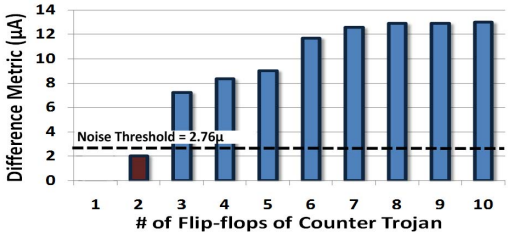


Fig. 6. Difference metric for varying sequential Trojan size in IEU circuit.

chip independently and do not compare the current signature between golden and infected ICs for Trojan detection.

The slight difference in current signatures for the original circuit is due to measurement noise, which is obtained from an FPGA-based experimental setup [8] and any difference larger than the computed noise threshold is attributed to the presence of a Trojan. The difference metric values for the test circuits with various Trojan instances are shown in Table I. The difference for a non-infected IC is also shown for comparison, which falls within the noise threshold. Table I also lists the test length obtained using our test vector generation tool which causes each rare node to go to its rare value $N = 20$ times. It should be noted that the entire test set is not needed for detecting counter- or LFSR-type Trojans.

Next, we insert different sizes of the counter Trojan in the IEU circuit to estimate the sensitivity of the approach. We changed the size from 10 bits to 1 bit and the corresponding values of the difference metric are plotted in Fig. 6. Even though TeSR fails to detect Trojans which have less than 2 flip-flops, such counters will activate their malicious payload in 4 cycles and can be detected using logic testing approaches. Moreover, measurement noise can be reduced further with better instrumentation (usually available in production test setup) to detect such Trojans by TeSR approach.

V. CONCLUSION

We have presented TeSR, a novel temporal self-referencing based side-channel analysis approach for hardware Trojan detection with high detection sensitivity. It facilitates detection of small, rarely-activated sequential Trojans, which can be extremely difficult to detect using existing logic testing or side-channel approaches. The approach leverages on the uncorrelated temporal variations in transient current signature of sequential hardware Trojans to isolate their effect from process and measurement noise. To the best of our knowledge, this is the first side-channel analysis approach for Trojan detection that 1) completely mitigates the effect of inter-die and intra-die process noise (both random and systematic); and 2) avoids the need to have golden reference chips, which may be difficult or highly expensive to obtain. The simulation results show that the proposed method can be very effective in isolating chips with hard-to-detect sequential Trojans of varying size, which can easily evade logic testing and other side-channel approaches, under large process noise.

REFERENCES

[1] M. Tehranipoor and F. Koushanfar, "A survey of hardware Trojan taxonomy and detection," *IEEE Design and Test of Computers*, 2010.

[2] R.S. Chakraborty, *et al*, "MERO: A statistical approach for hardware Trojan detection", *CHES Workshop*, 2009.

[3] L. Lin, W. Burleson and C. Parr, "MOLES: malicious off-chip leakage enabled by side-channels", *ICCAD*, 2009.

[4] D. Agrawal, *et al*, "Trojan detection using IC fingerprinting", *IEEE Symp. on Security and Privacy*, 2007.

[5] Y. Jin and Y. Makris, "Hardware Trojan detection using path delay fingerprint", *HOST*, 2008.

[6] S. Borkar, *et al*, "Parameter variations and impact on circuits and micro-architecture", *DAC*, 2003.

[7] R. Rad, J. Plusquellic and M. Tehranipoor, "A sensitivity analysis of power signal methods for detecting hardware Trojans under real process and environmental conditions", *IEEE TVLSI*, 2010.

[8] S. Narasimhan *et al*, "Multiple-parameter side-channel analysis: A non-invasive hardware Trojan detection approach", *HOST*, 2010.

[9] D. Du, *et al*, "Self-referencing: A scalable side-channel approach for hardware Trojan detection", *CHES*, 2010.

[10] M. Banga and M. Hsiao, "A region based approach for the identification of Hardware Trojans", *HOST*, 2008.

[11] M. Potkonjak, A. Nahapetian, M. Nelson and T. Massey, "Hardware Trojan horse detection using gate-level characterization", *DAC*, 2009.

[12] Predictive Technology Model, [Online] http://www.eas.asu.edu/∼ptm/