The Eighteenth International Conference on Advances in Circuits, Electronics and Micro-electronics
CENICS 2025
October 26, 2025 to October 30, 2025 - Barcelona, Spain

KOREA UNIVERSITY

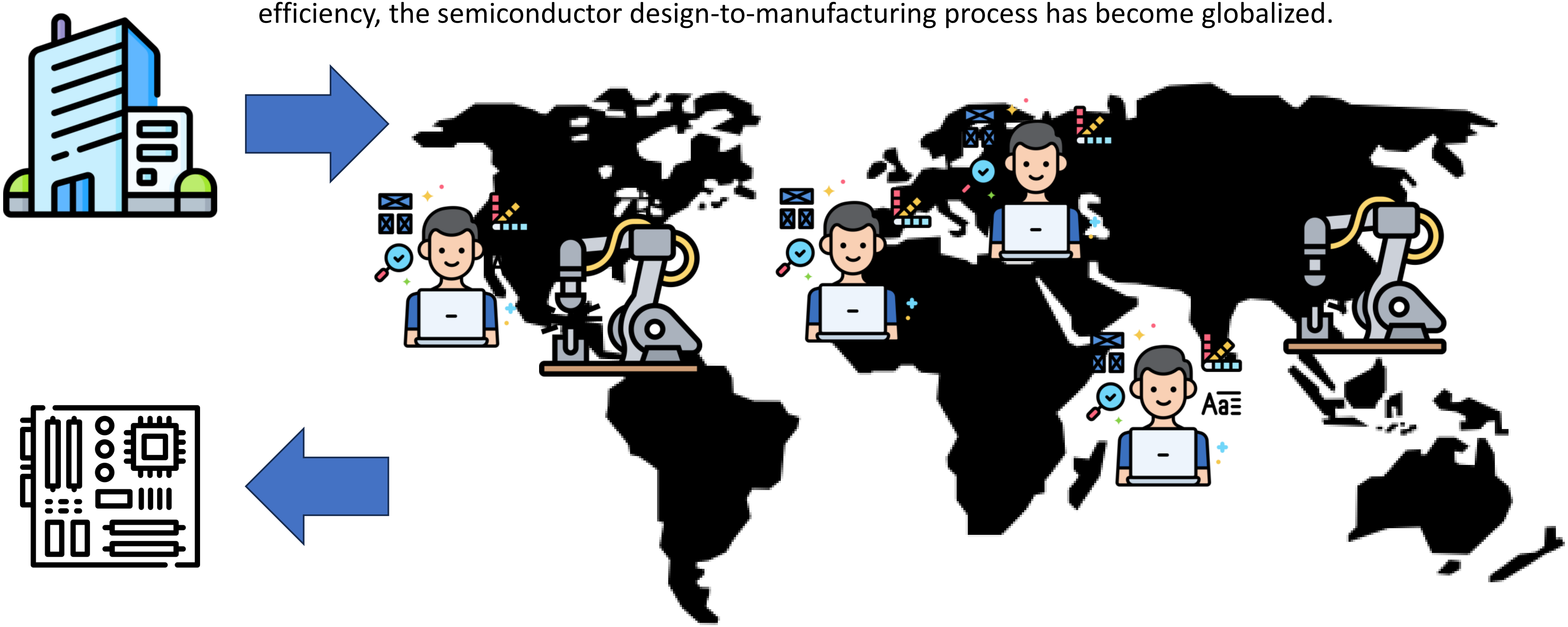
Malicous Hardware Detection with Advanced Artificial Intelligence Models

Junghee Lee

School of Cybersecurity, Korea University

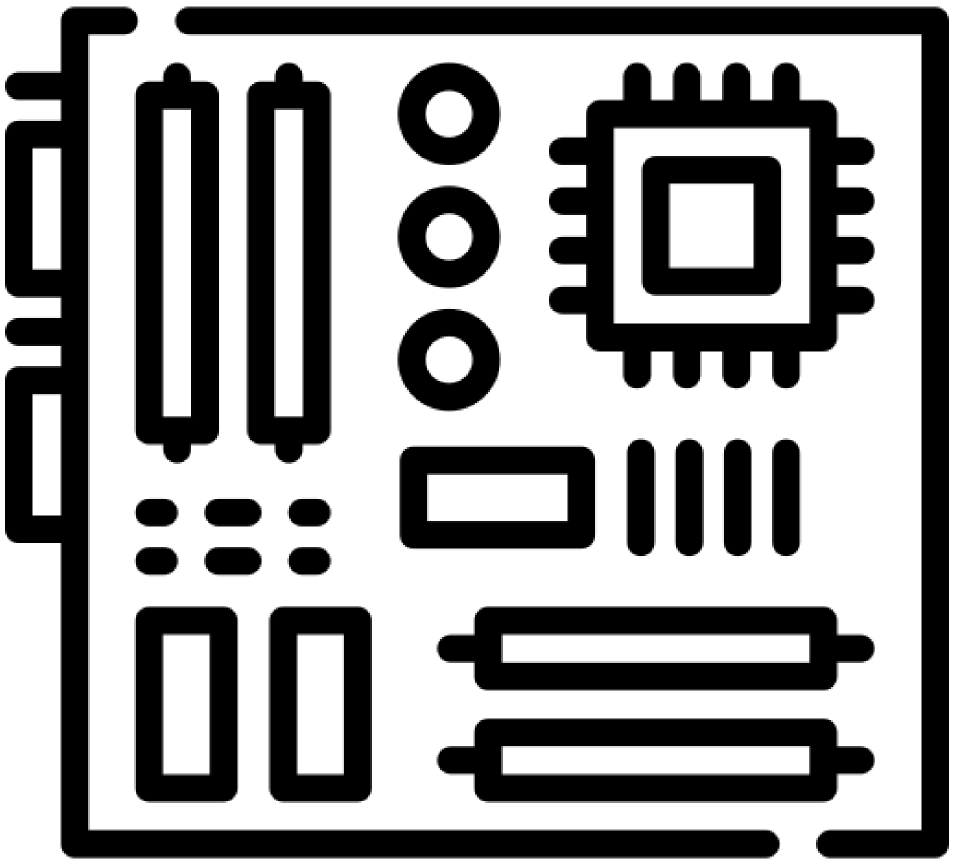
Background & Related works

Due to increasing technological complexity, cost-effectiveness, specialization, and supply chain efficiency, the semiconductor design-to-manufacturing process has become globalized.

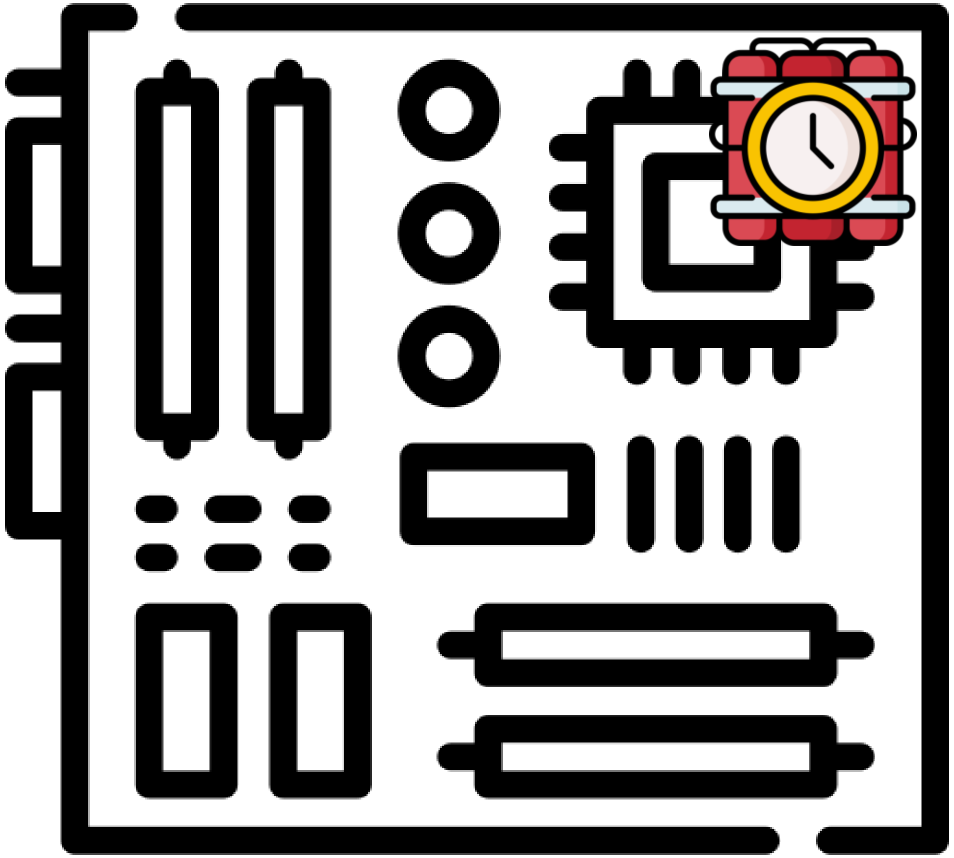


Background & Related works

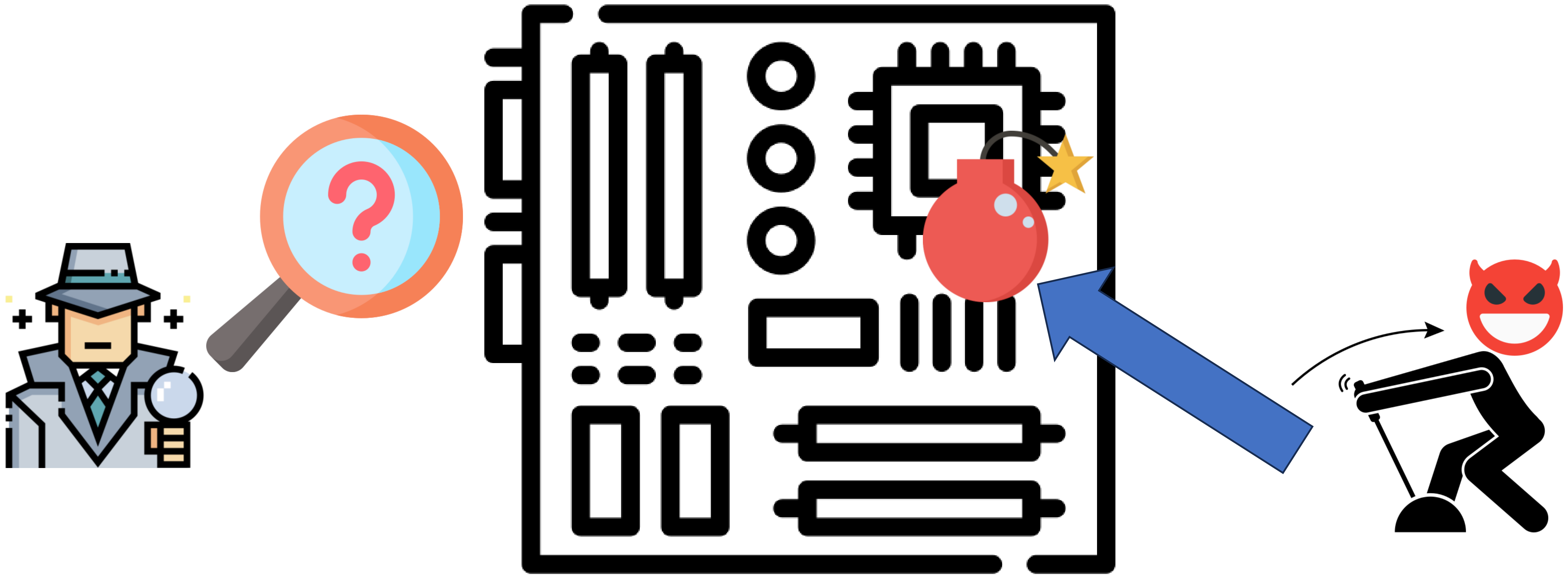
Do you think this hardware is secure?



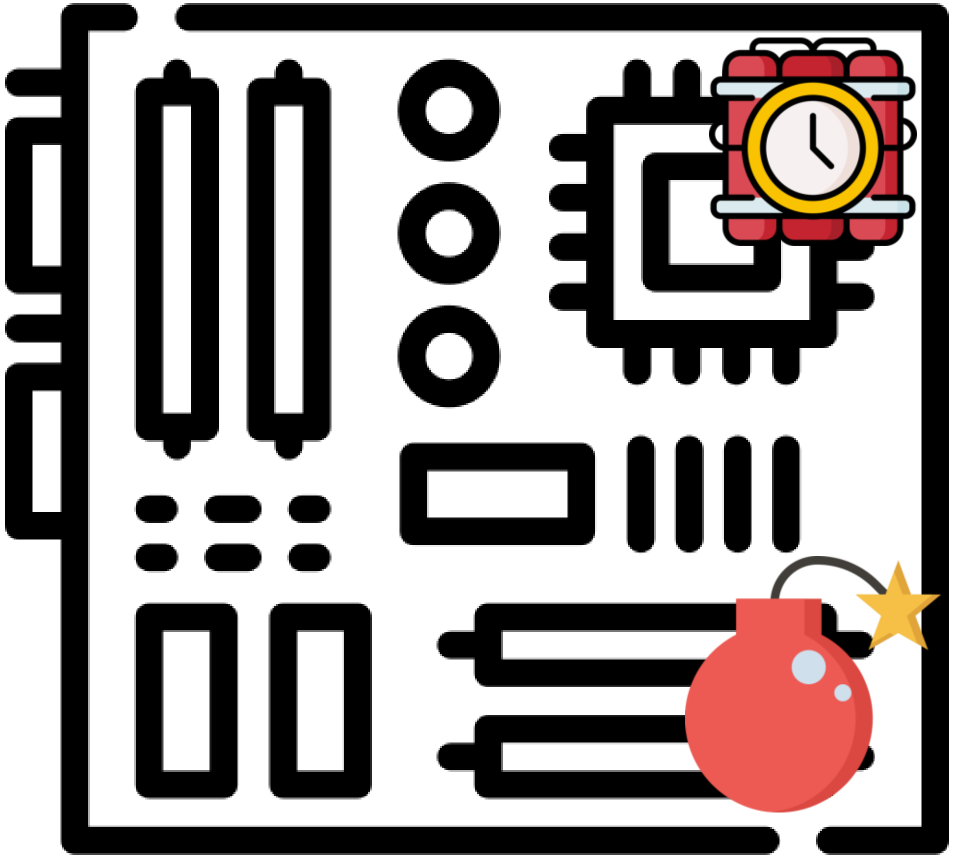
Background & Related works



Background & Related works



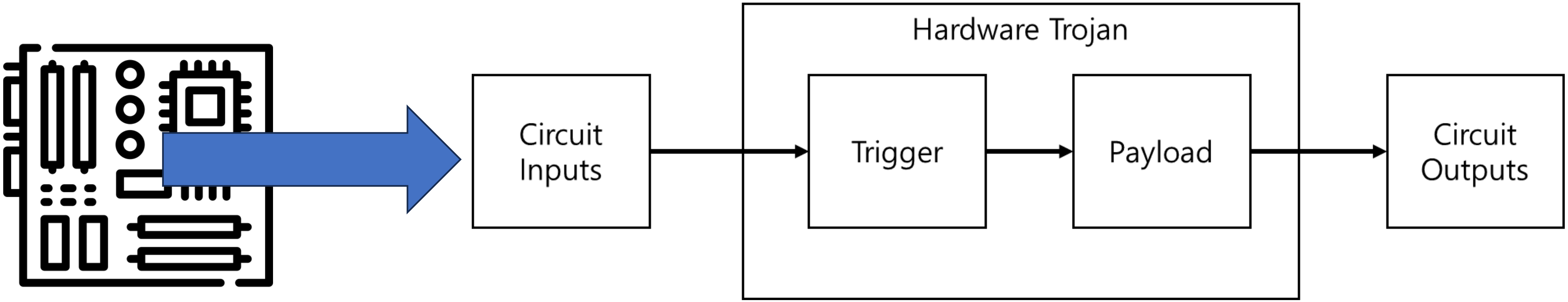
Background & Related works



This is Hardware Trojan!

Background & Related works

A Hardware Trojan(HT) refers to a malicious circuit component intentionally inserted during the design or manufacturing process of an integrated circuit (IC).



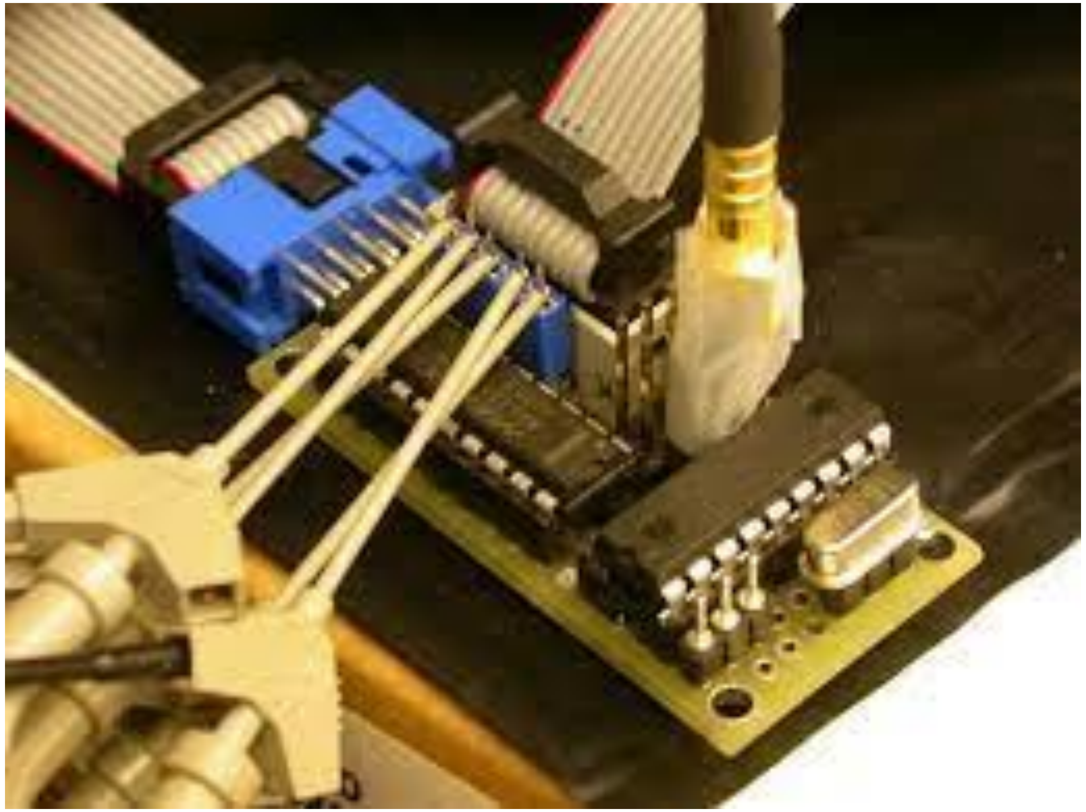
Potential Impacts of Hardware Trojans

- Information Leakage : HTs can leak sensitive data such as cryptographic keys, passwords, or biometric information to external entities.
- Denial of Service : HTs may halt system operation or cause abnormal behavior when certain conditions are met.
- Change functionality : An IC may appear to function normally, but HTs can trigger unintended behavior.
- Downgrade performance : changing power characteristics or introducing reliability problems in the chip. This influences power and/or delay characteristics of wires and gates in the affected circuit.

Background & Related works

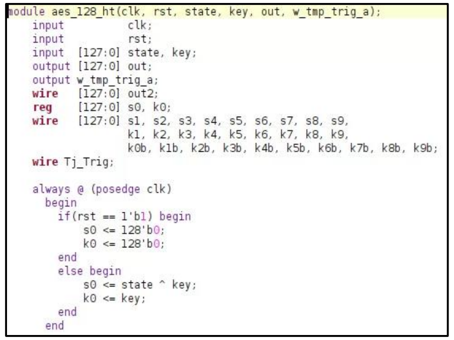
"Breakthrough silicon scanning discovers backdoor in military chip."

"The Big Hack"



Background & Related works

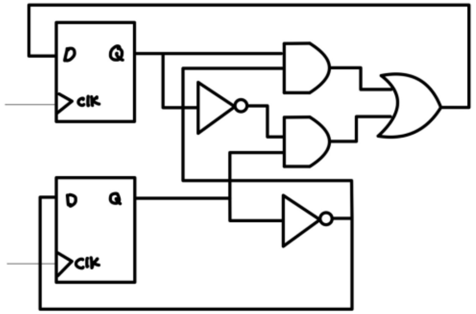
Specification

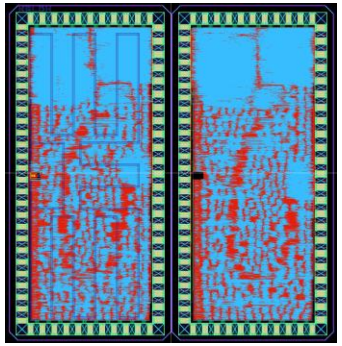


```
module aes_128_ht(clk, rst, state, key, out, w_tmp_trig_a);
    input           clk;
    input           rst;
    input   [127:0] state, key;
    output  [127:0] out;
    output w_tmp_trig_a;
    wire    [127:0] out2;
    reg     [127:0] s0, k0;
    wire    [127:0] s1, s2, s3, s4, s5, s6, s7, s8, s9,
                    k1, k2, k3, k4, k5, k6, k7, k8, k9,
                    k0b, k1b, k2b, k3b, k4b, k5b, k6b, k7b, k8b, k9b;
    wire Tj_Trig;

    always @ (posedge clk)
      begin
        if(rst == 1'b1) begin
            s0 <= 128'b0;
            k0 <= 128'b0;
        end
        else begin
            s0 <= state ^ key;
            k0 <= key;
        end
      end
```

RTL Design

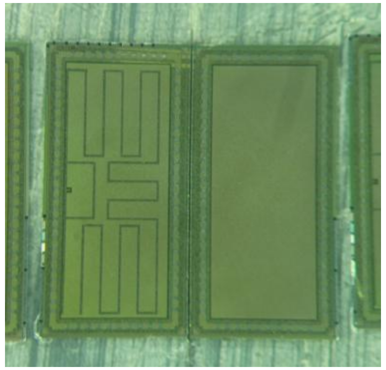
Netlist/Gate Level



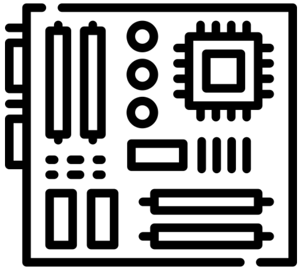


Physical Design Layout

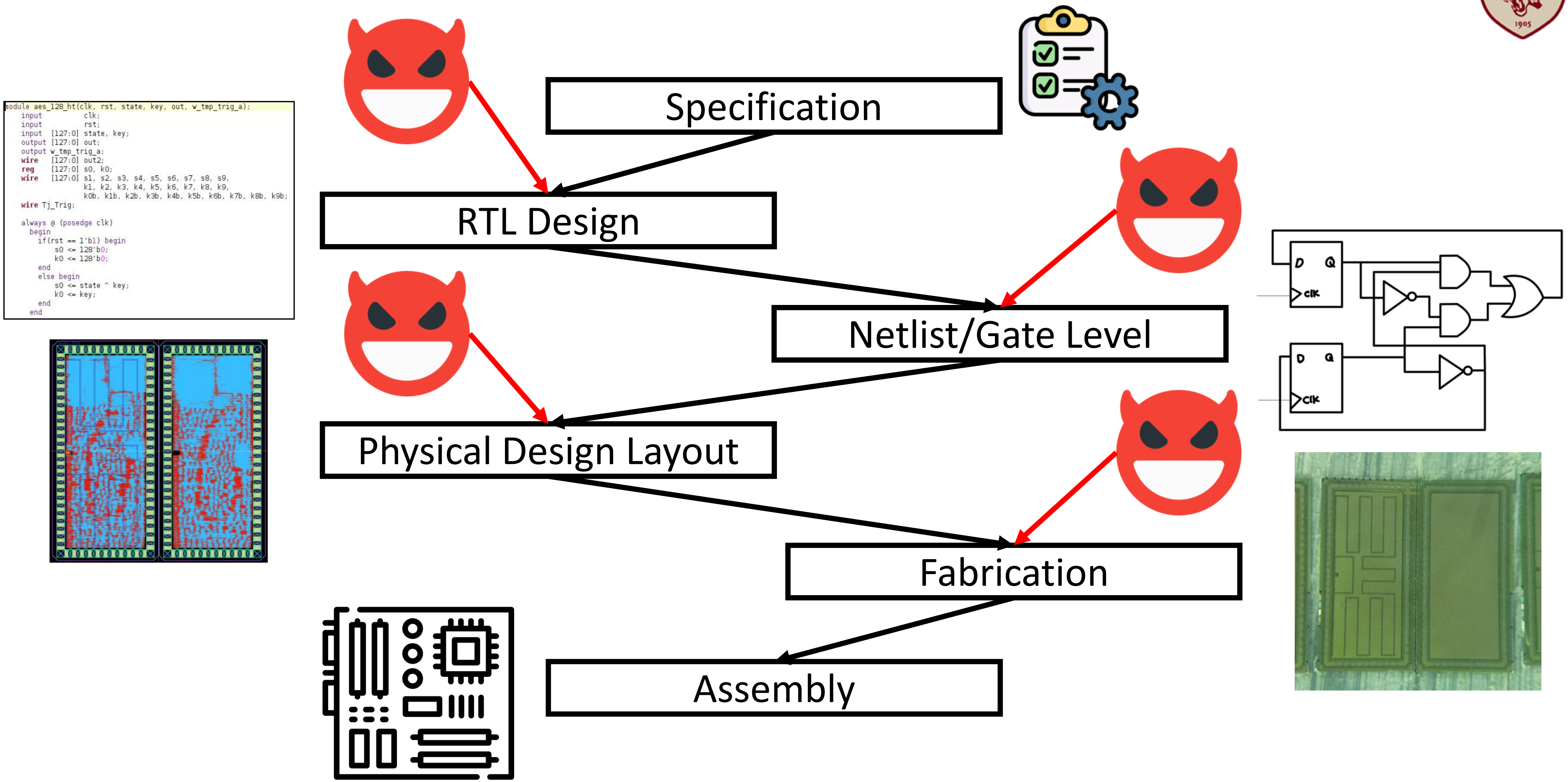
Fabrication



Assembly

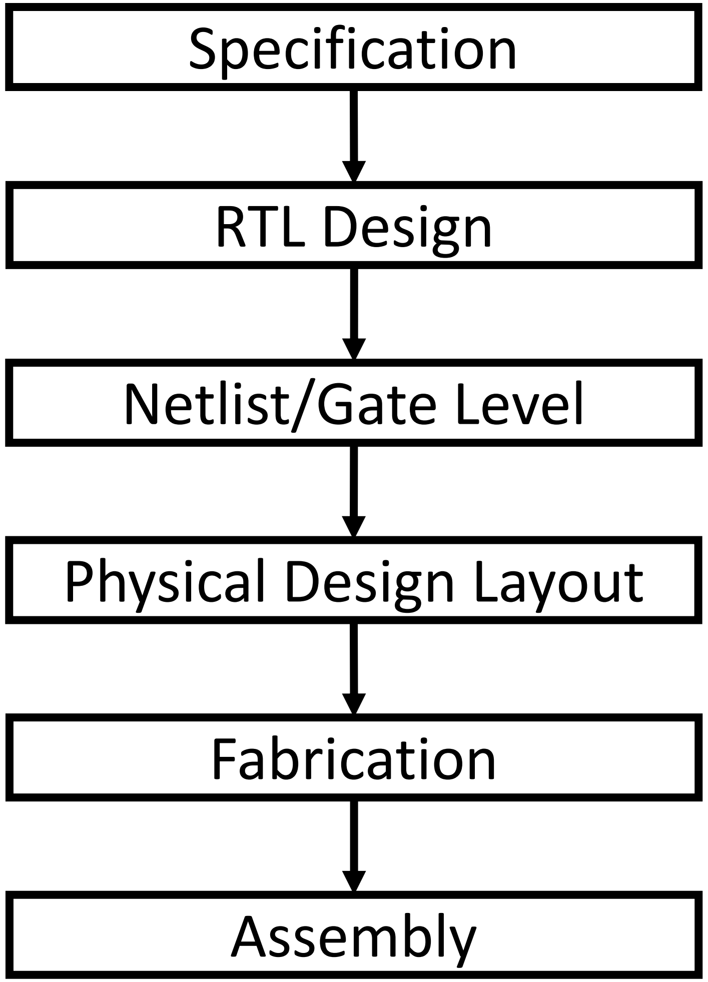


Background & Related works




```
module aes_128_ht(clk, rst, state, key, out, w_tmp_trig_a);
    input           clk;
    input           rst;
    input  [127:0] state, key;
    output [127:0] out;
    output w_tmp_trig_a;
    wire   [127:0] out2;
    reg    [127:0] s0, k0;
    wire   [127:0] s1, s2, s3, s4, s5, s6, s7, s8, s9,
                   k1, k2, k3, k4, k5, k6, k7, k8, k9,
                   k0b, k1b, k2b, k3b, k4b, k5b, k6b, k7b, k8b, k9b;
    wire Tj_Trig;

    always @ (posedge clk)
      begin
        if(rst == 1'b1) begin
            s0 <= 128'b0;
            k0 <= 128'b0;
        end
        else begin
            s0 <= state ^ key;
            k0 <= key;
        end
      end
```

Background & Related works



Why Is It Important to Detect Hardware Trojans?

- Prevent Security Breaches
  - ✓ HTs can act as critical vulnerabilities that compromise the integrity and confidentiality of entire systems.
  - ✓ They pose especially severe threats in domains like IoT, defense, autonomous vehicles, and financial systems, where a breach can cause catastrophic damage.
- Ensure Trustworthiness
  - ✓ It is essential to prove that chips manufactured in the product supply chain are secure.
  - ✓ HT detection is a prerequisite for gaining trust from customers and certification authorities.
- Detection Complexity
  - ✓ Since HTs are typically small in size and activate only under specific conditions, post-silicon validation (detection after manufacturing) is extremely challenging and costly.
  - ✓ Therefore, pre-silicon detection (during the design phase) is a much more effective and practical approach., and defense systems.

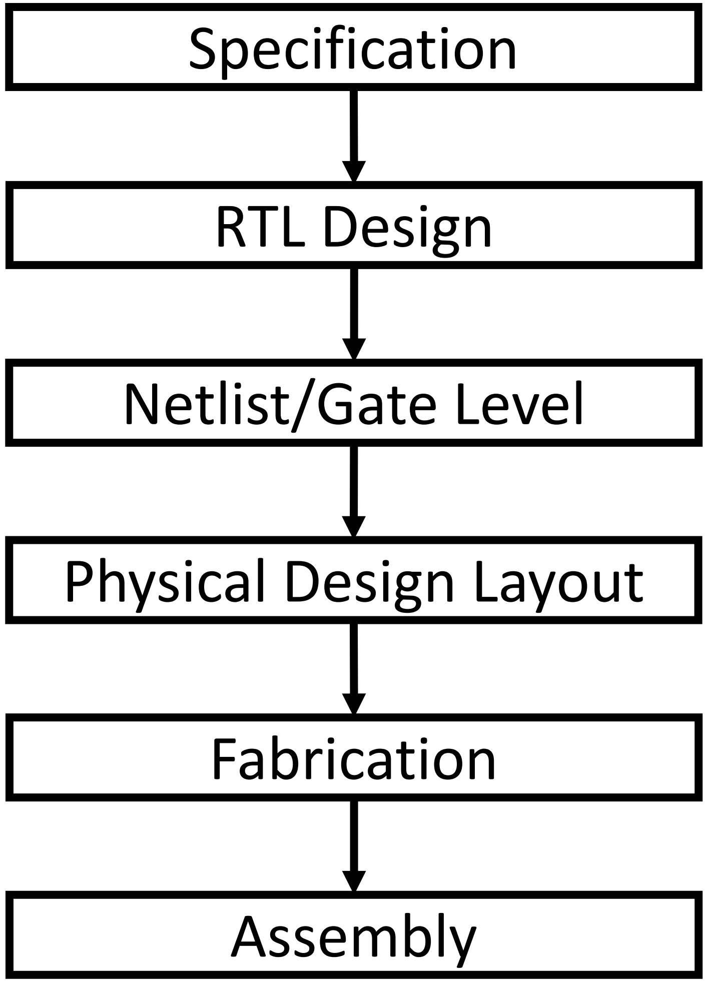
Background & Related works

Specification
↓
RTL Design
↓
Netlist/Gate Level
↓
Physical Design Layout
↓
Fabrication
↓
Assembly

Why Is It Important to Detect Hardware Trojans?

- **Prevent Security Breaches**
  - ✓ **HTs can act as critical vulnerabilities that compromise the integrity and confidentiality of entire systems.**
  - ✓ **They pose especially severe threats in domains like IoT, defense, autonomous vehicles, and financial systems, where a breach can cause catastrophic damage.**
- Ensure Trustworthiness
  - ✓ It is essential to prove that chips manufactured in the product supply chain are secure.
  - ✓ HT detection is a prerequisite for gaining trust from customers and certification authorities.
- Detection Complexity
  - ✓ Since HTs are typically small in size and activate only under specific conditions, post-silicon validation (detection after manufacturing) is extremely challenging and costly.
  - ✓ Therefore, pre-silicon detection (during the design phase) is a much more effective and practical approach., and defense systems.

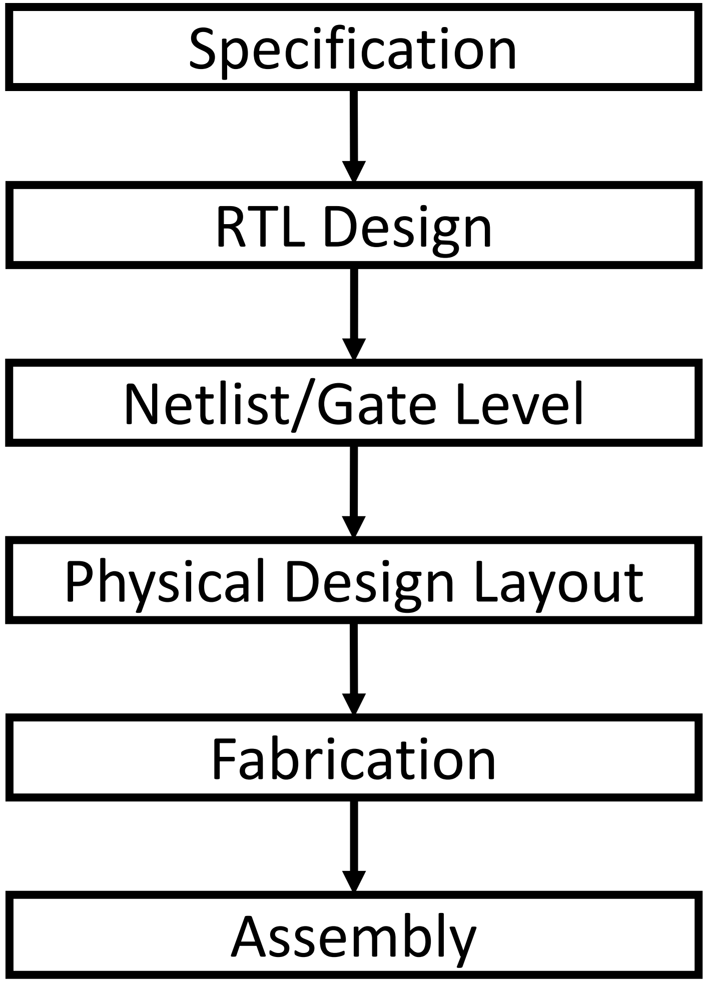
Background & Related works



Why Is It Important to Detect Hardware Trojans?

- Prevent Security Breaches
  - ✓ HTs can act as critical vulnerabilities that compromise the integrity and confidentiality of entire systems.
  - ✓ They pose especially severe threats in domains like IoT, defense, autonomous vehicles, and financial systems, where a breach can cause catastrophic damage.
- **Ensure Trustworthiness**
  - ✓ **It is essential to prove that chips manufactured in the product supply chain are secure.**
  - ✓ **HT detection is a prerequisite for gaining trust from customers and certification authorities.**
- Detection Complexity
  - ✓ Since HTs are typically small in size and activate only under specific conditions, post-silicon validation (detection after manufacturing) is extremely challenging and costly.
  - ✓ Therefore, pre-silicon detection (during the design phase) is a much more effective and practical approach., and defense systems.

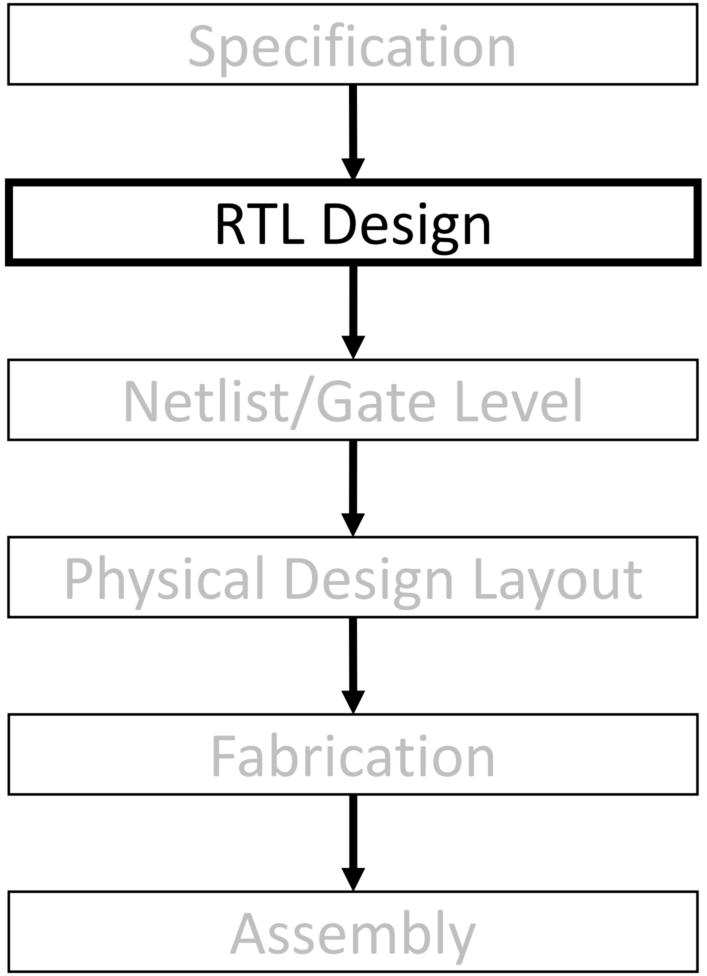
Background & Related works



Why Is It Important to Detect Hardware Trojans?

- Prevent Security Breaches
  - ✓ HTs can act as critical vulnerabilities that compromise the integrity and confidentiality of entire systems.
  - ✓ They pose especially severe threats in domains like IoT, defense, autonomous vehicles, and financial systems, where a breach can cause catastrophic damage.
- Ensure Trustworthiness
  - ✓ It is essential to prove that chips manufactured in the product supply chain are secure.
  - ✓ HT detection is a prerequisite for gaining trust from customers and certification authorities.
- **Detection Complexity**
  - ✓ **Since HTs are typically small in size and activate only under specific conditions, post-silicon validation (detection after manufacturing) is extremely challenging and costly.**
  - ✓ **Therefore, pre-silicon detection (during the design phase) is a much more effective and practical approach., and defense systems.**

Background & Related works



Advantages of Detecting Hardware Trojans at the RTL Design Stage

- Pre-Silicon Security Verification
  - ✓ The RTL (Register Transfer Level) stage represents the logical design of a circuit before it is physically fabricated on silicon.
  - ✓ Detecting HTs at this stage allows security vulnerabilities to be eliminated early, significantly reducing the cost and time associated with redesign.
- Enhances Supply Chain Security
  - ✓ By using models that can verify the absence of HTs during the design stage, it becomes possible to ensure higher trustworthiness in the subsequent manufacturing and testing phases.
  - ✓ This contributes to strengthening the overall security of the IC supply chain.

Background & Related works

Specification
↓
RTL Design
↓
Netlist/Gate Level
↓
Physical Design Layout
↓
Fabrication
↓
Assembly

Advantages of Detecting Hardware Trojans at the RTL Design Stage

- **Pre-Silicon Security Verification**
  - ✓ **The RTL (Register Transfer Level) stage represents the logical design of a circuit before it is physically fabricated on silicon.**
  - ✓ **Detecting HTs at this stage allows security vulnerabilities to be eliminated early, significantly reducing the cost and time associated with redesign.**
- Enhances Supply Chain Security
  - ✓ By using models that can verify the absence of HTs during the design stage, it becomes possible to ensure higher trustworthiness in the subsequent manufacturing and testing phases.
  - ✓ This contributes to strengthening the overall security of the IC supply chain.

Background & Related works

Specification → **RTL Design** → Netlist/Gate Level → Physical Design Layout → Fabrication → Assembly

Advantages of Detecting Hardware Trojans at the RTL Design Stage

- Pre-Silicon Security Verification
  - ✓ The RTL (Register Transfer Level) stage represents the logical design of a circuit before it is physically fabricated on silicon.
  - ✓ Detecting HTs at this stage allows security vulnerabilities to be eliminated early, significantly reducing the cost and time associated with redesign.
- **Enhances Supply Chain Security**
  - ✓ **By using models that can verify the absence of HTs during the design stage, it becomes possible to ensure higher trustworthiness in the subsequent manufacturing and testing phases.**
  - ✓ **This contributes to strengthening the overall security of the IC supply chain.**

Proposed method

Our Goal

Can we be certain that there are no HTs if the RTL is designed according to the designer’s intent?

Proposed method

Our Goal

Can we be certain that there are no HTs if the RTL is designed according to the designer’s intent?

However, in reality, during the integration process after RTL design, the use of third-party IPs, or the RTL utilization stage, an attacker can stealthily insert HTs without disrupting the intended functionality. Therefore, functional verification alone is insufficient to detect HTs, and it is necessary to identify structural anomalies within circuits that appear functionally normal.

Proposed method

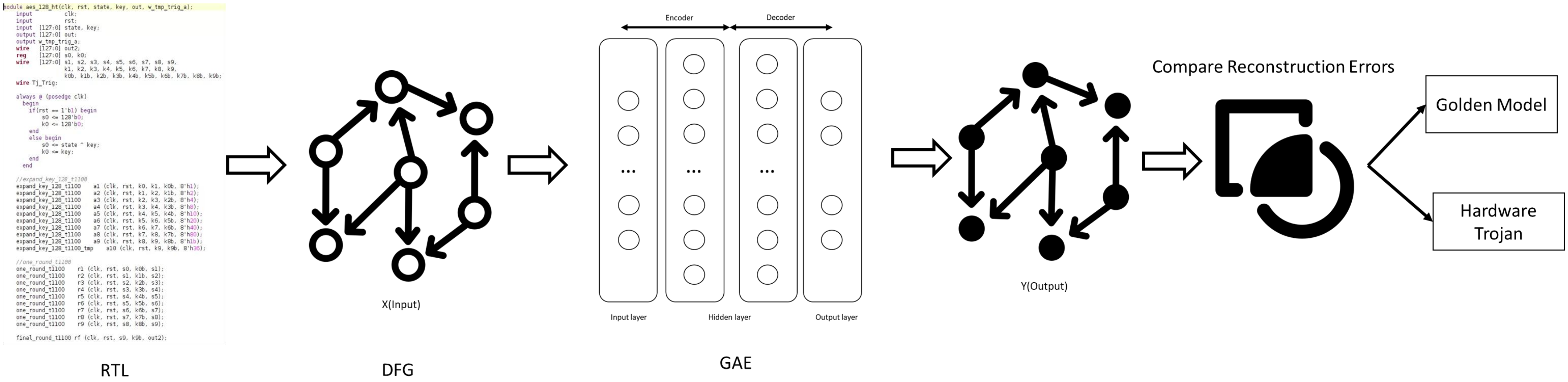
Our Goal

We proposes a method to verify, at the design stage, the absence of HTs by detecting structurally anomalous regions—despite functional normality—using an unsupervised learning-based Graph Autoencoder (GAE).

Proposed method

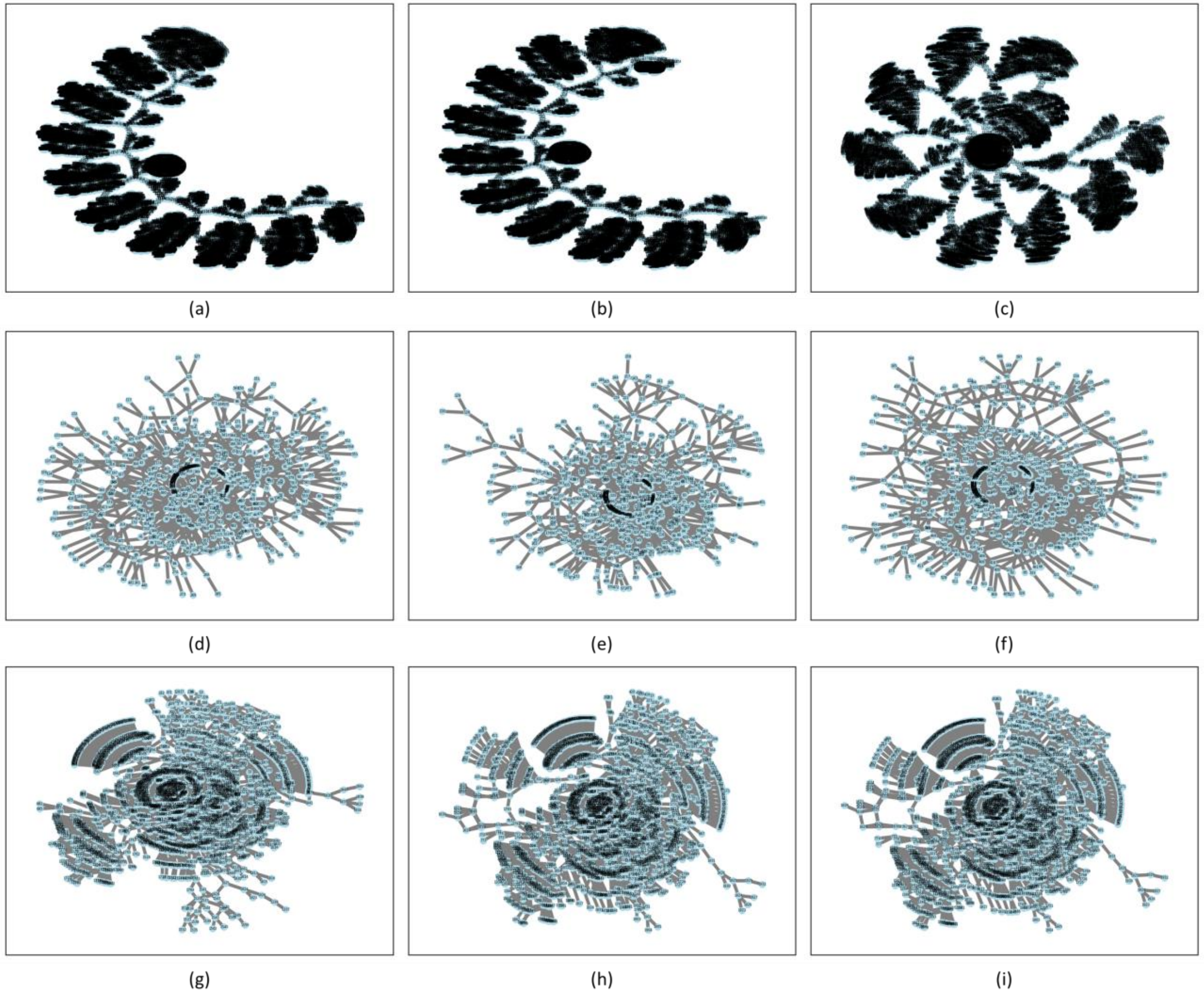
Our Goal

High-level a schematic of our method



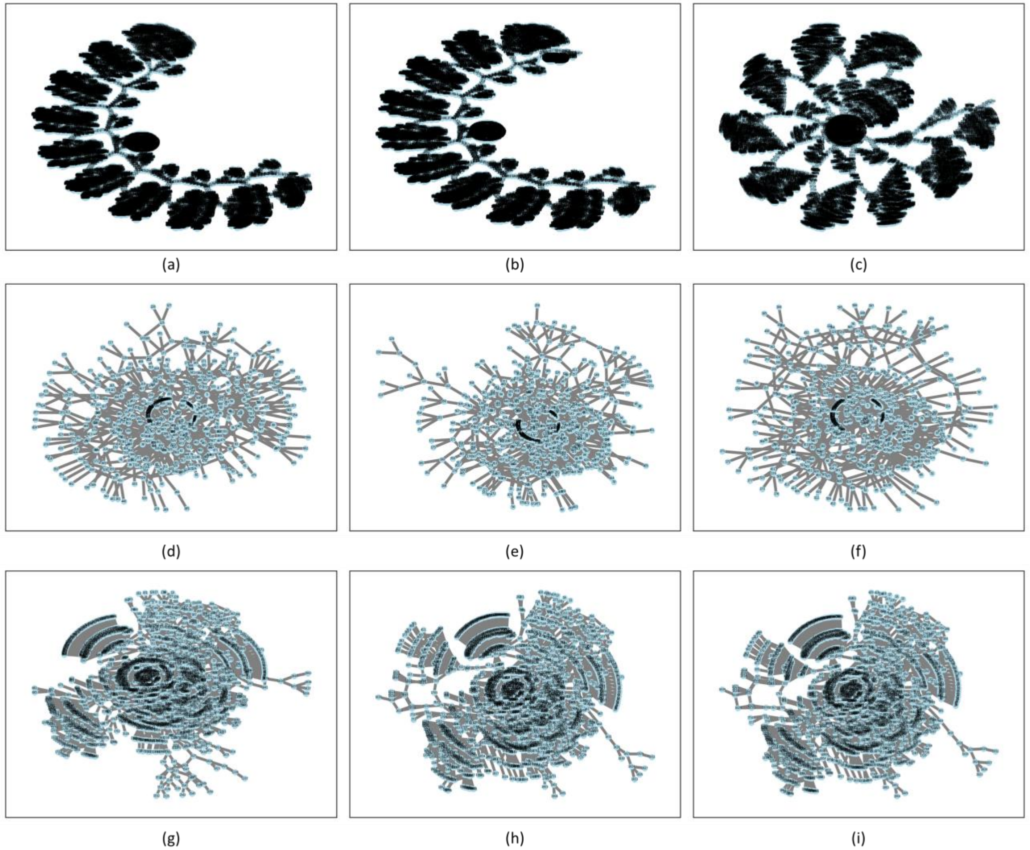
Proposed method

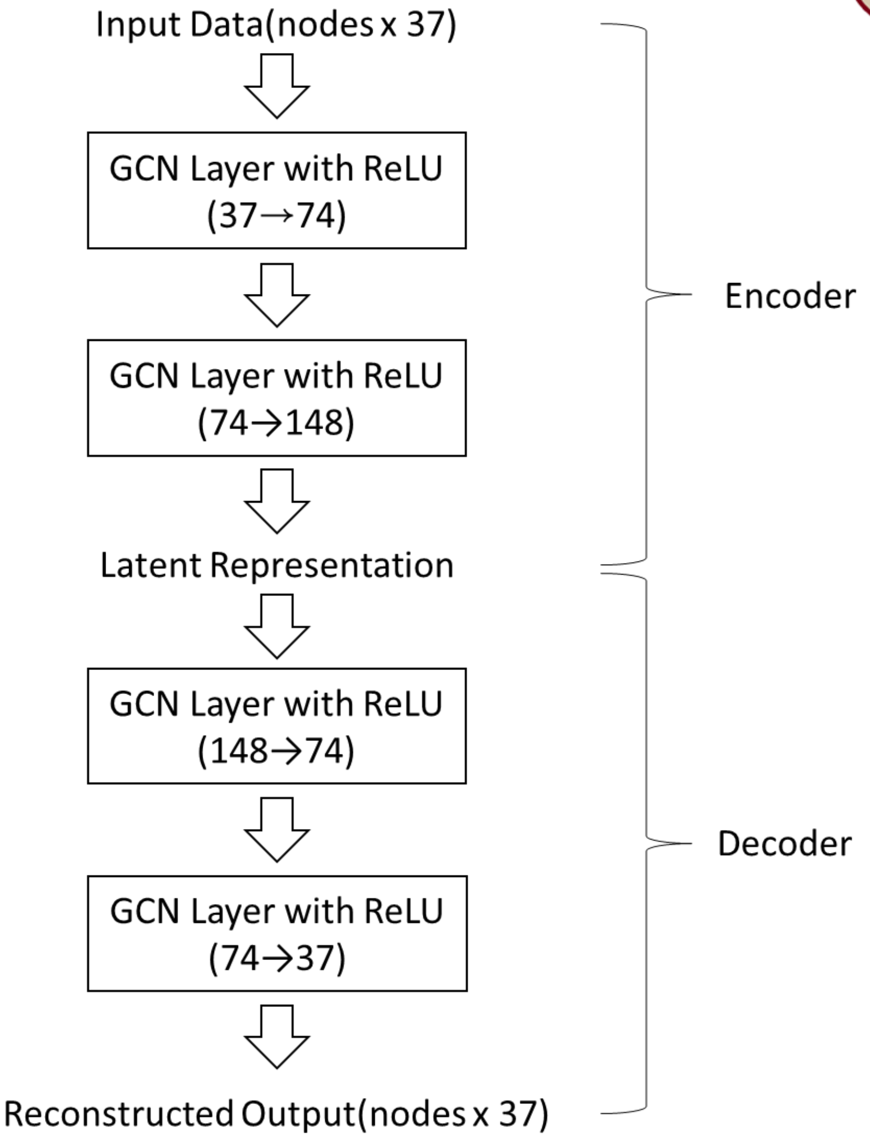
```
module aes_128_ht(clk, rst, state, key, out, w_tmp_trig_a);
    input           clk;
    input           rst;
    input   [127:0] state, key;
    output  [127:0] out;
    output w_tmp_trig_a;
    wire    [127:0] out2;
    reg     [127:0] s0, k0;
    wire    [127:0] s1, s2, s3, s4, s5, s6, s7, s8, s9,
                    k1, k2, k3, k4, k5, k6, k7, k8, k9,
                    k0b, k1b, k2b, k3b, k4b, k5b, k6b, k7b, k8b, k9b;
    wire Tj_Trig;

    always @ (posedge clk)
      begin
        if(rst == 1'b1) begin
            s0 <= 128'b0;
            k0 <= 128'b0;
        end
        else begin
            s0 <= state ^ key;
            k0 <= key;
        end
      end
```



Convert RTL(GM) to graph data

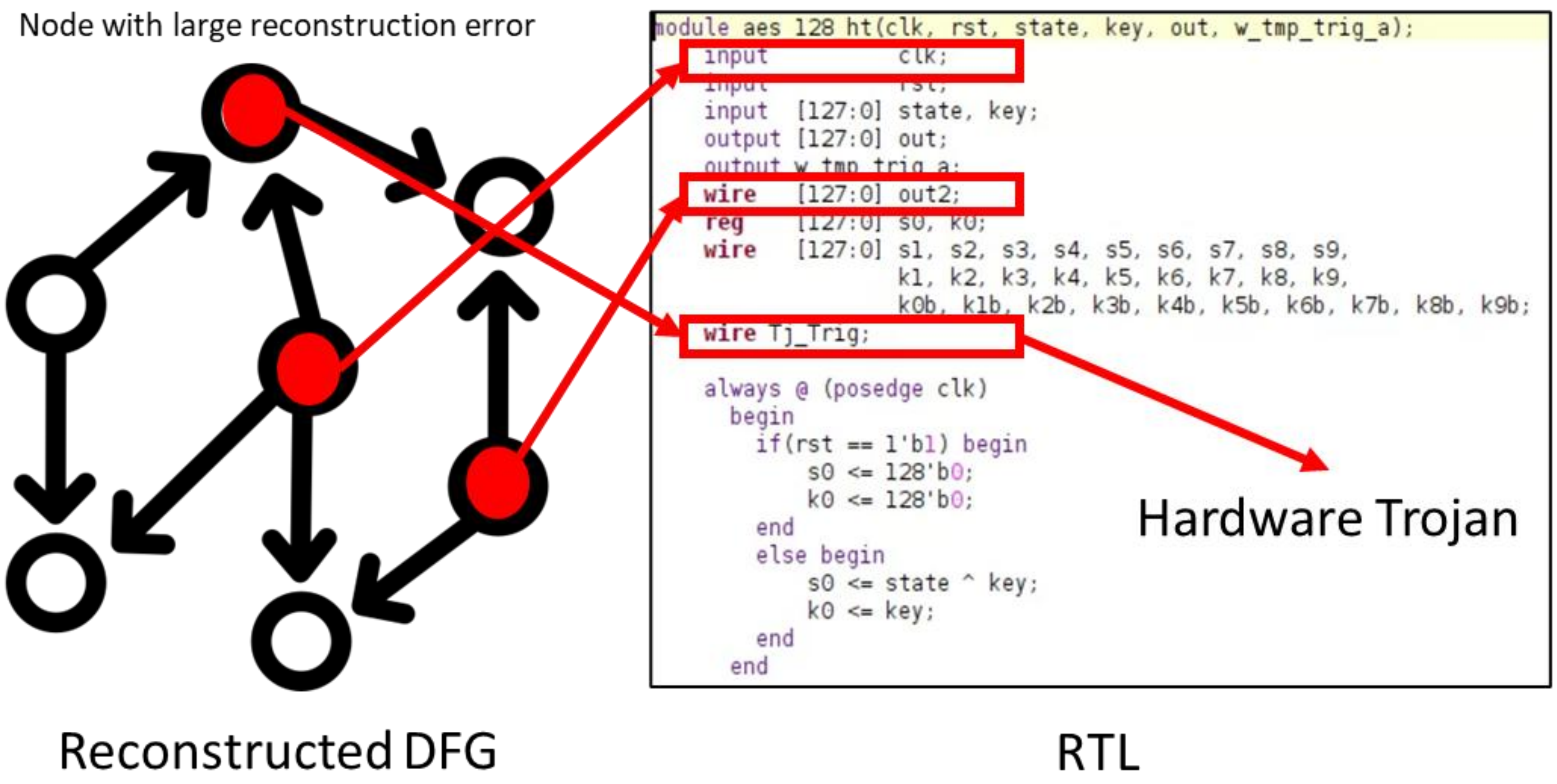
Proposed method





Using learning and test data with GAE

Proposed method



If the reconstruction error exceeds the threshold (Th), it is determined that an HT has been inserted, and by sorting the reconstruction errors of each node, the suspicious nodes can be mapped to the corresponding RTL code.

Results

Table 5: The modularity of untrained HTs and their similarity to the GM for performance comparison.

Benchmark	Detection difficulty metric		Successful HT detection				Localization accuracy (%)					
	Modularity	Similarity	[36]	[6]	[24]	Ours	$L(1)$	$L(2)$	$L(5)$	$L(10)$	$L(20)$	$L(All)$
AES-T200	0.971417564	0.999990410	O	O	O	O	100.0	100.0	60.00	80.00	85.71	81.82
AES-T300	0.969626044	0.999984572	O	O	O	O	100.0	100.0	100.0	60.00	60.00	70.37
AES-T400	0.973308165	0.999994401	X	O	O	O	100.0	100.0	100.0	80.00	89.47	88.24
AES-T500	0.972776842	0.999998631	X	O	O	O	100.0	100.0	80.00	80.00	88.89	83.33
AES-T600	0.970348372	0.999998286	X	O	O	O	100.0	100.0	80.00	90.00	95.00	90.91
AES-T700	0.972131557	0.999993565	O	O	O	O	100.0	100.0	60.00	80.00	86.67	80.00
AES-T800	0.971612477	0.999993409	O	O	O	O	100.0	100.0	100.0	80.00	90.00	87.50
AES-T900	0.972227887	0.999993553	O	O	O	O	100.0	100.0	60.00	80.00	87.50	81.82
AES-T1000	0.971606293	0.999988274	O	O	O	O	100.0	100.0	60.00	80.00	88.24	81.82
AES-T1100	0.974109880	0.999979723	O	O	O	O	100.0	100.0	100.0	80.00	90.00	89.47
AES-T1200	0.971335814	0.999990147	O	O	O	O	100.0	100.0	60.00	80.00	88.24	84.62
AES-T1300	0.976191388	0.999983454	O	O	O	O	100.0	100.0	100.0	50.00	70.00	77.14
AES-T1400	0.972557902	0.999984456	O	O	O	O	100.0	100.0	40.00	40.00	60.00	77.14
AES-T1500	0.976239173	0.999970516	O	X	O	O	100.0	100.0	40.00	30.00	60.00	70.97
AES-T1600	0.976982885	0.999967829	X	X	O	O	100.0	100.0	100.0	90.00	90.00	91.67
AES-T1700	0.975548388	0.999978077	X	X	O	O	100.0	100.0	100.0	80.00	87.50	85.71
AES-T1800	0.976065260	0.999989559	X	O	O	O	100.0	100.0	80.00	50.00	83.33	75.00
AES-T1900	0.976024699	0.999972474	X	X	O	O	100.0	100.0	80.00	50.00	83.33	75.00
AES-T2000	0.971457001	0.999985900	X	O	O	O	100.0	100.0	100.0	90.00	95.00	92.31
RS232-T200	0.822624646	0.999805259	O	O	O	O	100.0	50.00	40.00	50.00	70.00	57.89
RS232-T300	0.803968802	0.999879782	X	X	O	O	100.0	50.00	60.00	70.00	63.16	81.25
RS232-T400	0.815408721	0.999856644	X	X	O	O	100.0	50.00	60.00	60.00	78.57	66.67
RS232-T500	0.810271366	0.999863642	X	X	O	O	100.0	100.0	80.00	80.00	63.16	78.95
RS232-T600	0.792732479	0.999292380	X	X	O	O	100.0	100.0	80.00	60.00	80.00	75.00
RS232-T700	0.792732479	0.999292380	X	X	O	O	100.0	100.0	80.00	60.00	80.00	75.00
RS232-T800	0.798013320	0.999331041	O	O	O	O	100.0	50.00	80.00	70.00	70.00	58.33
RS232-T900	0.779747438	0.999287553	X	X	O	O	100.0	100.0	100.0	60.00	73.68	75.00
RS232-T901	0.785438256	0.999170823	X	X	O	O	100.0	100.0	100.0	80.00	75.00	80.00
PIC16F84-T200	0.833086809	0.999990499	O	O	O	O	100.0	50.00	20.00	10.00	50.00	7.41
PIC16F84-T300	0.833204190	0.999990499	O	O	O	O	100.0	50.00	20.00	10.00	50.00	7.69
PIC16F84-T400	0.833093852	0.999990499	O	O	O	O	100.0	50.00	20.00	10.00	50.00	8.33

Experimental Objective

- To verify whether the proposed unsupervised GAE model can detect various Hardware Trojans (HTs) without requiring labeled HT data.
- To evaluate whether structural anomalies in RTL designs with inserted Trojans can be identified after training only on the Golden Model (GM).

Results

- Accuracy : Comparable to or better than existing methods (100% Accuracy / F1-score)
- Generality : Capable of detecting HTs regardless of their type or insertion location
- Localization : Accurate identification of insertion points using only the top-1 reconstruction error node

Results

Table 5: The modularity of untrained HTs and their similarity to the GM for performance comparison.

Benchmark	Detection difficulty metric		Successful HT detection				Localization accuracy (%)					
	Modularity	Similarity	[36]	[6]	[24]	Ours	$L(1)$	$L(2)$	$L(5)$	$L(10)$	$L(20)$	$L(\text{All})$
AES-T200	0.971417564	0.999990410	O	O	O	O	100.0	100.0	60.00	80.00	85.71	81.82
AES-T300	0.969626044	0.999984572	O	O	O	O	100.0	100.0	100.0	60.00	60.00	70.37
AES-T400	0.973308165	0.999994401	X	O	O	O	100.0	100.0	100.0	80.00	89.47	88.24
AES-T500	0.972776842	0.999998631	X	O	O	O	100.0	100.0	80.00	80.00	88.89	83.33
AES-T600	0.970348372	0.999998286	X	O	O	O	100.0	100.0	80.00	90.00	95.00	90.91
AES-T700	0.972131557	0.999993565	O	O	O	O	100.0	100.0	60.00	80.00	86.67	80.00
AES-T800	0.971612477	0.999993409	O	O	O	O	100.0	100.0	100.0	80.00	90.00	87.50
AES-T900	0.972227887	0.999993553	O	O	O	O	100.0	100.0	60.00	80.00	87.50	81.82
AES-T1000	0.971606293	0.999988274	O	O	O	O	100.0	100.0	60.00	80.00	88.24	81.82
AES-T1100	0.974109880	0.999979723	O	O	O	O	100.0	100.0	100.0	80.00	90.00	89.47
AES-T1200	0.971335814	0.999990147	O	O	O	O	100.0	100.0	60.00	80.00	88.24	84.62
AES-T1300	0.976191388	0.999983454	O	O	O	O	100.0	100.0	100.0	50.00	70.00	77.14
AES-T1400	0.972557902	0.999984456	O	O	O	O	100.0	100.0	40.00	40.00	60.00	77.14
AES-T1500	0.976239173	0.999970516	O	X	O	O	100.0	100.0	40.00	30.00	60.00	70.97
AES-T1600	0.976982885	0.999967829	X	X	O	O	100.0	100.0	100.0	90.00	90.00	91.67
AES-T1700	0.975548388	0.999978077	X	X	O	O	100.0	100.0	100.0	80.00	87.50	85.71
AES-T1800	0.976065260	0.999989559	X	O	O	O	100.0	100.0	80.00	50.00	83.33	75.00
AES-T1900	0.976024699	0.999972474	X	X	O	O	100.0	100.0	80.00	50.00	83.33	75.00
AES-T2000	0.971457001	0.999985900	X	O	O	O	100.0	100.0	100.0	90.00	95.00	92.31
RS232-T200	0.822624646	0.999805259	O	O	O	O	100.0	50.00	40.00	50.00	70.00	57.89
RS232-T300	0.803968802	0.999879782	X	X	O	O	100.0	50.00	60.00	70.00	63.16	81.25
RS232-T400	0.815408721	0.999856644	X	X	O	O	100.0	50.00	60.00	60.00	78.57	66.67
RS232-T500	0.810271366	0.999863642	X	X	O	O	100.0	100.0	80.00	80.00	63.16	78.95
RS232-T600	0.792732479	0.999292380	X	X	O	O	100.0	100.0	80.00	60.00	80.00	75.00
RS232-T700	0.792732479	0.999292380	X	X	O	O	100.0	100.0	80.00	60.00	80.00	75.00
RS232-T800	0.798013320	0.999331041	O	O	O	O	100.0	50.00	80.00	70.00	70.00	58.33
RS232-T900	0.779747438	0.992287553	X	X	O	O	100.0	100.0	100.0	60.00	73.68	75.00
RS232-T901	0.785438256	0.991170823	X	X	O	O	100.0	100.0	100.0	80.00	75.00	80.00
PIC16F84-T200	0.833086809	0.999990499	O	O	O	O	100.0	50.00	20.00	10.00	50.00	7.41
PIC16F84-T300	0.833204190	0.999990499	O	O	O	O	100.0	50.00	20.00	10.00	50.00	7.69
PIC16F84-T400	0.833093852	0.999990499	O	O	O	O	100.0	50.00	20.00	10.00	50.00	8.33

Evaluation of HT Localization Performance

- L(n) metric: The percentage of the top n nodes with the highest reconstruction errors that correspond to actual HT locations.
- L(1): Always 100% → The node with the highest reconstruction error exactly matches the HT insertion point.
- Although HT localization becomes more challenging in highly complex circuits, the accuracy based on L(1) consistently remains at 100%.

Results

Table 6: The top 10 list of high reconstruction errors when the AES128-T300 data flow graph is reconstructed.

Node index	Reconstruction errors	Node name
11280	2.985195875	top.TjTrojan1.SHReg7
11275	2.985195398	top.TjTrojan1.rk4
11276	2.985195398	top.TjTrojan1.rk5
11277	2.985195398	top.TjTrojan1.rk6
11278	2.985195398	top.TjTrojan1.rk7
11279	2.985195398	top.TjTrojan1.rk8
11229	1.706307411	top.AES.rf.stateout
2558	1.684933305	top.AES.r2.clk
3583	1.684933305	top.AES.r3.clk
4608	1.684933305	top.AES.r4.clk

Table 7: The top 20 list of high reconstruction errors with GM and HTs. One node from each HT is highlighted.

2 HTs (GM+AES128-T200+AES128-T400) $L(20)$=45%	3 HTs (2 HTs + AES128-T500) $L(20)$=55%	4 HTs (3 HTs + AES128-T600) $L(20)$=55%
top.TjTSC400.Tjbeep2	top.TjTSC400.Tjbeep2	top.TjTSC400.Tjbeep2
top.TjTSC400.TjSHIFTReg	top.TjTSC400.TjSHIFTReg	top.TjTSC400.TjSHIFTReg
top.TjTSC400.Tjbeeps	top.TjTSC400.Tjbeeps	top.TjTSC400.Tjbeeps
top.AES.key	top.AES.key	top.AES.key
top.TjTrojan200.Tjlfsr.Tjlfsrstream	top.Capacitance	top.Capacitance
top.Capacitance	top.TjTrojan200.Tjlfsr.Tjlfsrstream	top.TjTrojan200.Tjlfsr.Tjlfsrstream
top.AES.clk	top.TjTSC400.TjMUXSel	top.AES.clk
top.TjTSC400.TjTrig	top.AES.clk	top.TjTSC400.TjTrig
top.AES.rf.S41.S1.out	top.TjTSC400.TjTrig	top.TjTrojan600.TjDetected
top.AES.rf.S42.S1.out	top.AES.rf.S41.S1.out	top.AES.rf.S41.S1.out
top.AES.rf.S43.S1.out	top.AES.rf.S42.S1.out	top.AES.rf.S42.S1.out
top.AES.rf.S44.S1.out	top.AES.rf.S43.S1.out	top.AES.rf.S43.S1.out
top.TjTrojan400.TjDetected	top.AES.rf.S44.S1.out	top.AES.rf.S44.S1.out
top.TjTSC400.Tjrst	top.TjTrojan400.TjDetected	top.TjTrojan400.TjDetected
top.TjTrojan400.Tjclk	top.TjTrojan500.TjDetected	top.TjTSC400.Tjrst
top.AES.state	top.TjTSC400.Tjrst	top.TjTrojan500.TjState2
top.Antena	top.TjTrojan400.Tjclk	top.TjTrojan500.TjState3
top.TjTSC400.Tjkey	top.TjTrojan500.Tjclk	top.TjTrojan400.Tjclk
top.AES.r1.t1.t3.s0.out	top.AES.state	top.AES.state
top.AES.r1.t2.t3.s0.out	top.Antena	top.Antena

Evaluation of HT Localization Performance

- The top 10 nodes with the highest reconstruction errors were confirmed to correspond to actual HT circuits.
- Even when 2, 3, or 4 HTs were simultaneously inserted into AES128, all were successfully detected.
- Among the top 20 nodes with the highest reconstruction errors, nodes corresponding to each HT were included.
- At least one node from every HT appeared in the top-ranked list.

Discussion & Conclusions

Limitations

- The method proposed in this study requires GM.
- Since the GAE is specialized and trained on a single GM, its generalization performance to other circuits (GMs) is limited.
- Additional research is needed to detect a wider variety of HT types.

Contributions

- GAE4HT demonstrated very high performance in both HT detection and localization(100% Accuracy and F1-score), effectively overcoming the limitations of traditional supervised learning and side-channel-based methods.
- It requires no HT data for training; instead, it performs unsupervised learning solely based on the GM, enabling effective detection and localization of HTs across various types and complexity levels.
- This study presents a practical and scalable solution that facilitates early-stage HT detection during the IC design and verification process, helping to preemptively block potential security threats.

Discussion & Conclusions

“Assume every chip is unusual”

Thank you