Automating LLMs for Hardware Trojan Insertion -- Automating Trojan Definitions

Objective

Students will expand upon the previous hardware Trojan assignments by using generative AI to aid not only with generating additional Trojans, but also with identifying where those Trojans should be added and what kind of Trojans.

Tasks

You must generate 15 unique Trojans in the OpenTitan SoC (5 denial of service, 5 information leakage, 5 change of functionality).

The Trojan generation must be fully automated using either GHOST or another tool of your choosing.

Along with generating the Trojans, you must also use generative AI to analyze the design and find vulnerable locations for the Trojans. This part does not need to be fully automated.

OpenTitan

OpenTitan is a fully open-source silicon root-of-trust, and will be the target of your Trojans for this assignment. A silicon root-of-trust like OpenTitan is a component which contains security primitives and other security features that you can use as the base level of trustworthiness in its connected hardware. As such, it makes for an incredibly high-value target for potential adversaries.

As with the previous challenges, you must be able to show that OpenTitan still functions (simulates) as expected in their normal testing apparatus. You must also be able to provide all of your logs/conversations/etc. from using AI for these tasks.

You can find information on OpenTitan at their official website as well as their GitHub

OpenTitan provides their own official Docker container which has all of the tools the project needs. We **highly** recommend you make use of this Docker container for this assignment, as OpenTitan relies on several tools which can be complex to set up on your own.

Submission

Due 15 December
You must include the following in your GitHub repo:

- A report containing:
    - Explanations of how each Trojan works
    - How you tested each Trojan
    - Any troubleshooting steps/design decisions you needed to take with the automated system
- Logs of all LLM interactions (we need to be able to see what you prompted and what the model(s) produced)
- All modified RTL
- Testbenches used to check the Trojans

Organize your repo in the following manner:

```
My_Repo
├── Report
├── trojan_1
│   ├── rtl/
│   │   └── <all modified RTL>
│   ├── tb/
│   │   └── <testbench to exploit Trojan>
│   └── ai/
│       └── <all AI interactions (chat logs, etc.)>
├── trojan_2
│   ├── rtl/
│   │   └── <all modified RTL>
│   ├── tb/
│   │   └── <testbench to exploit Trojan>
│   └── ai/
│       └── <all AI interactions (chat logs, etc.)>
├── etc ...
```