netidee

STIPENDIEN

Attack Detection using Micro-architectural Traces and Machine Learning

Zwischenbericht | Call 19 | Stipendium ID 7276

Contents

1 Introduction

This project focuses on detecting side-channel attacks including Spectre, Rowhammer and Zombieload by analyzing microarchitectural traces from hardware components. By collecting data from hardware performance counters (HPCs) and applying machine learning techniques, we aim to develop an efficient and accurate detection framework. The goal is to classify running workloads as benign or malicious based on subtle patterns in the hardware-level behavior, helping to enhance system security against advanced threats.

2 Status

2.1 **Meilenstein 1 -** <Spectre Attack Detection>

Over the past period, I finalized and conducted experiments on detecting Spectre attacks. I also published a research paper detailing these results (Statistical Profiling of Micro-Architectural Traces and Machine Learning for Spectre Detection: A Systematic Evaluation) and gave presentations within my research group and partner institutions. These activities were focused on sharing findings and refining our approach through feedback and collaboration.

In recent years, microarchitecture side-channel attacks have emerged as a serious threat, not because they directly damage hardware like the CPU, but because they exploit subtle hardware behaviors to leak sensitive information. These attacks operate quietly in the background, leaving no visible trace for users but exposing secrets such as encryption keys, passwords, or other confidential data. Although they don't disrupt normal system operations, they leave behind subtle footprints in microarchitectural components, such as cache usage patterns or branch predictor behavior which can be monitored.

To address this, we developed a framework that leverages hardware-level performance monitoring tools to collect microarchitectural traces during program execution. These traces, which can capture the indirect effects of side-channel activity, are then analyzed using machine learning models to distinguish between benign and malicious workloads. By systematically experimenting with various combinations of features and classifiers, we identified the most effective configurations for accurate detection. Our initial findings, including evaluations across diverse scenarios, have been published in our first research paper.

Our results showed that using only four hardware performance counters (HPCs) are Conditional Branch Instructions Executed (BR_CN), Conditional Branch Instructions Corr. Pred (BR_PRC), Total Branch Instructions Executed (BR_INS), and Total Cycles Executed (TOT_CYC) are sufficient for detecting Spectre attacks with high accuracy. The top-performing classifiers were Decision Tree (DT), Gradient Boosting Classifier (GBC), and Random Forest (RF). Our model effectively classifies workloads as benign or malicious based on these HPCs.

2.2 **Milestone 2 -** <Writing a publication about Spectre Attack Detection >:

We achieved strong results and published our first paper at a well-regarded conference (DATE Conference : Design, Automation and Test in Europe Conference | The European Event for Electronic System Design & Test)

The main challenges were at the hardware level and involved technical limitations. However, these were successfully addressed through the use of alternative tools and continuous experimentation. Fortunately, no major deviations occurred, and all activities are progressing according to plan.

3 Summary of plan update

We have successfully completed Milestone 1, which focused on detecting Spectre attacks using microarchitectural traces and machine learning. As planned, we have now moved on to Milestone 2, which involves expanding the detection framework to include additional types of side-channel attacks. There have been no major adjustments to the planning document, as this progression aligns with the original project timeline. Our current work includes modeling and evaluating new detection models for a broader range of attacks. We also aim to publish a second paper presenting these extended results and findings.