Research Summary

I am Tinghao (Vitus) Xie, a junior undergraduate student majoring in Computer Science and Technology at Zhejiang University (ZJU). My current research interest lies in the intersection of secure, efficient, robust AI and systems.

I've been guided by Prof. *Jianhai Chen* in ZJU since 2018, both in **Super Computing Team (SCT)** and the **Intelligent Computing and System (INCAS)** Lab. I've recently started another research internship around AI security advised by Prof. *Shouling Ji* and *Xuhong Zhang* in **NESA Lab**.

ZJU Super Computing Team (ZJUSCT)

I learned about high-performance computing in SCT, where I also gained the certificate of competency of Accelerated computing basics – CUDA C/C++ from Nvidia Deep Learning Institute. As a member of SCT, I participated in ASC Student Supercomputer Challenge 2020-2021 and did the analyzation and optimization work surrounding QuEST, a quantum circuit simulator.

INCAS Lab

Since May 2020, I've joined the INCAS lab and have been conducting research around **system security** (*SGX Security Protection Technology of Distributed Machine Learning under GPU Architecture*). I designed *Enchecap*, a simple encrypted heterogeneous computation protocol and am currently implementing it with Intel SGX and CUDA on untrusted servers.

ALPS Lab

I am now a novice intern conducting research about **backdoor attack and defense** at ALPS Lab, PSU, advised by Professor Ting Wang.

All the research work I have done is shown on this page.

Enchecap: An **enc**rypted (**enc**lave-based) **he**terogeneous **ca**lculation **p**rotocol based on Nvidia CUDA and Intel SGX

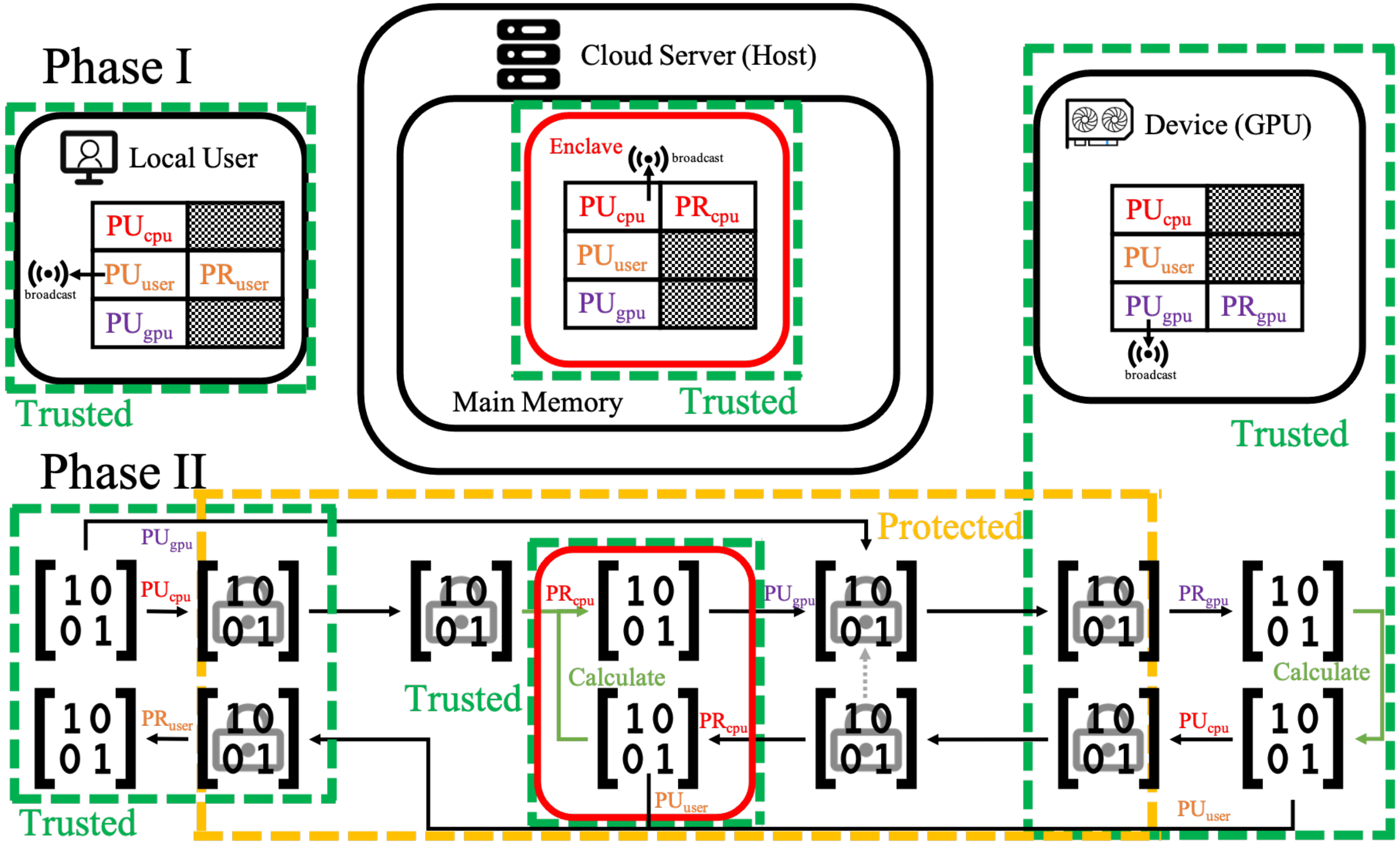
Enchecap is a system-level protocol for trusted heterogeneous computation, based on Intel SGX (a hardware TEE technology). The protocol is implemented on Nvidia GPU and with Nvidia CUDA toolkit.

Background

Cloud computing is widely applied nowadays, including sensitive-data driven tasks. Also, heterogeneous computation with accelerators like GPU is increasingly necessary. Under such circumstances, those critical data in computation might be easily stolen by cloud administrators or any attackers with such a privilege. Thus, protections at the system level are desperately needed.

Threat Model

We are assuming an adversary with administrator privileges and physical access to the cloud server host memory. However, the adversary can only peek but not overwrite. Also, the protection of the cloud device (GPU) is out of range; we assume there are some identity authentication features (works like Graviton could provide) when accessing the GPU memory.



Enchecap. The protocol ensures data security **during network transmission**, **on main memory**, and **at host-device I/Os** (under the threat model described earlier). Local user, SGX enclave and GPU could be regarded as 3 trusted agent. In Phase I, they broadcast and record public keys (PU). In Phase II, calculation could happen both in the enclave and on GPU. data would be decrypted and visible only in trusted area, and encrypted before going out.

Contribution

- We propose **Enchecap**, an **enc**rypted (**enc**lave-based) **he**terogeneous **ca**lculation **p**rotocol. It provides certain defenses against the threat model described earlier.
- We implement the protocol into a library wrapping up related functions for handy deployments, and demonstrate the protocol practically with the CUBLAS sample. The overhead attached to a single round trip in a 2-matrix multiplication CUBLAS sample of is around 38%.

Some of my projects are shown on this page (also available through my CV).

Research on the Texture Packing Problem

A group research project focusing on approximation algorithms solving the texture(strip) packing problem, a 2D version of the bin packing problem, report available here.

Contribution

- Conducted research on different texture(strip) packing algorithms
- Combined the genetic algorithm with traditional approximation algorithms
- Analyzed performance of various algorithm combinations

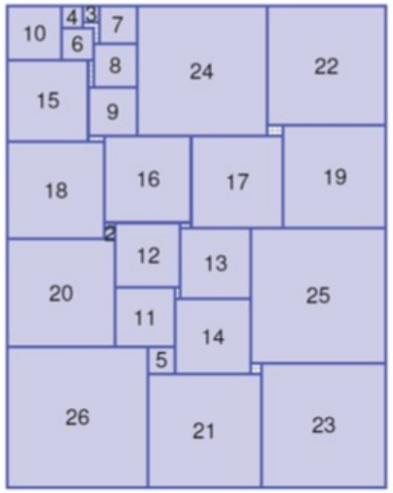
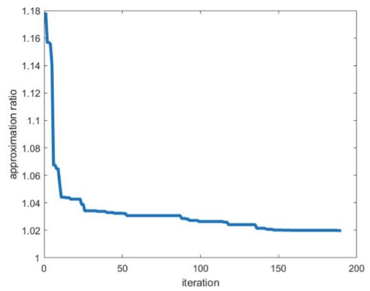
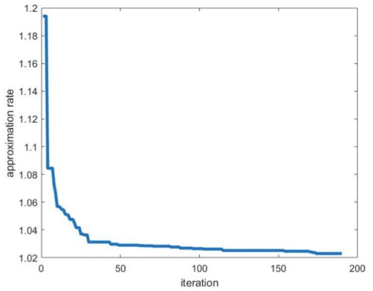
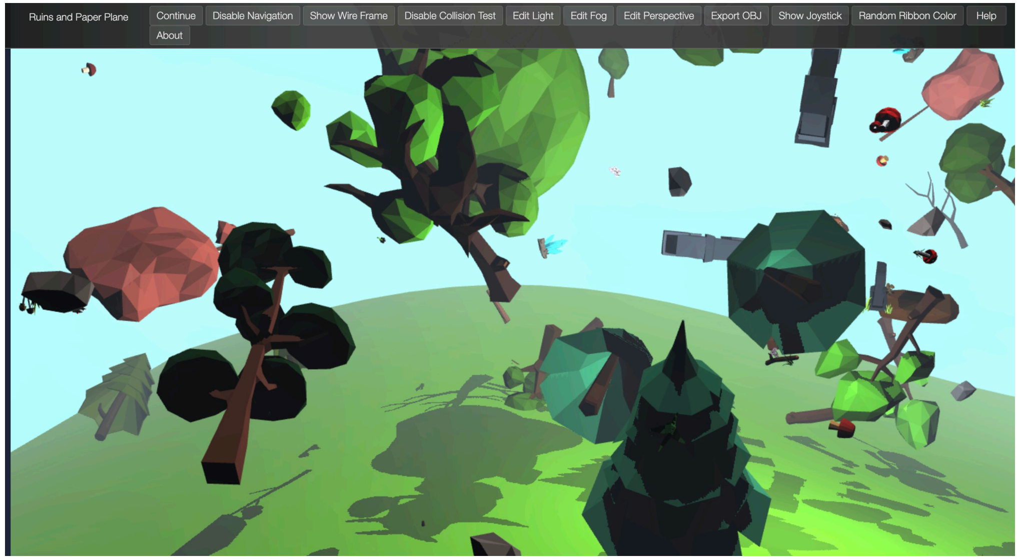


Figure 1: An example for *Texture Packing* problem







Tron: A 3D Graphic Engine Based on WebGL

Tron is a rendering engine based on native WebGL with a wonderful flying game demo.

Contribution

- Designed the representation pattern and data structures for 3D scenes
- Completed voxel, material and texture expression modules
- Wrote GLSL shader codes involving fogs and the animated sky
- Implemented cross-platform interaction and front-end web pages