Dahua Feng

Website / GitHub wwh8us@virginia.edu

Education

B.S. in Information and Computing Sciences, Peking University Beijing, China
School of Electronics Engineering and Computer Science Sep. 2020 - Jul. 2024

- **GPA:** 3.700/4.000 (87.4/100)
- **Thesis:** *A Simulator Design for the Storage of Mobile Devices* (supervised by Prof. Jie Zhang)

Publications

1. **Profiling Apple Silicon Performance for ML Training**, Preprint, 2025
   Dahua Feng*, Zhiming Xu*, Rongxiang Wang, Felix Xiaozhu Lin

Research Experience

School of Computer Science, Peking University Mar. 2022 - Jan. 2023
Research Intern Advisor: Prof. Zhi Yang
Project: DNN acceleration based on graph optimization

- This project focused on the acceleration for neural network computation. Primarily, we tried to use the genetic algorithm based on BFS and DP to schedule the ops. Then we mainly focused on the resources allocation for each of the operations.
- Analyzed part of the source code of Roller and the time evaluation source code of TVM.
- Implemented the functionality of resources allocation for the IOS scheduler and obtained about 10% improvement.

Picasso Lab, University of California, Santa Barbara Jul. 2023 - Dec. 2023
Research Intern Advisor: Prof. Yufei Ding
Project: CXL-based memory disaggregated system for DLRM

- This project aimed to improve the performance of the DLRM (deep learning recommendation system) training. We tried to implement a better approach for sharding embedding tables across many GPUs with CXL as memory expansion units.
- Analyzed the problem of the embedding table placement on multi-GPU theoretically and proposed some possible algorithms for the load-balance memory allocation.
- Explored the existing approach to solving sharding problems such as using RL and proposed some potential ways to improve it.

Project: Combination of NVSHMEM with DLRM embedding table lookup

- This project is about to leveraging the unbalanced placement of embedding tables by introducing NVSHMEM as a new inter-GPU communication method. We used CUDA kernel to get realistic data to help us design the strategy.
- Analyzed the realistic data and got some insights about the influence of placement.
- Completed some important components such as a prediction model to predict the time of table batched embedding.

XSEL Lab, University of Virginia Aug. 2024-Dec. 2024
Research Assistant Advisor: Prof. Felix Lin
Project: Apple Silicon for machine learning

- This project focuses on the performance of ML training on Apple Silicon, targeting the affordable and democratized ML training. Profiled and analyzed the performance gap between Apple Silicon and NVIDIA GPUs in ML training, especially for LLM training.
- Conducted the experiments about end-to-end training performance of some SOTA generative models and benchmarked BLAS kernels as micro-benchmark, on various of hardward platforms.

- Organized the data and made related analysis based on profiling results, as well as providing the recommendation for practitioners.

SEPT Lab, University of Southern California Jun. 2025-Present

Research Assistant Advisor: Prof. Mengyuan Li

Project: Communication overhead in a GPU confidential computing environment

- This project focuses on the performance impact due to the confidential computing(CC) during ML inference or training. Current CC authentication method is lack of parallelism and has been a bottleneck.
- Proposed a scheme based on the block index to realize the parallelization, as well as lazy decryption mechanism to solve the problem of decryption and re-encryption of unused data .
- Carried on the background survey, implemented the whole system and conducted the evaluation.

Teaching Experience

Teaching Assistant of Computer Architectures Fall 2023

School of Electronic Engineering and Computer Science, Peking University

- As a TA of Computer Architectures course instructed by Prof. Jie Zhang, participated in the designing and revision of the course projects, conducted Q&A class sessions, organized the quizzes, and worked on the final exam.

Skills

Programming Languages & Softwares: C, C++, Python, CUDA C
Python Packages: torch, scikit-learn
Languages: Mandarin (native), English (TOEFL iBT: 102/120), Korean (TOPIK 228/300, Level-5)