

Dahua Feng

School of Electronics Engineering and
Computer Science, Peking University
5 Yiheyuan Road, Haidian District, Beijing

(+86)18811391895
fdh23187@stu.pku.edu.cn
[GitHub](#)

Education

B.S. in Information and Computing Sciences, Peking University

Beijing, China

School of Electronics Engineering and Computer Science

Sep. 2020 - Jul. 2024 Expected

- **GPA:** 3.653/4.000 **Rank:** 25/96

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.
- I analyzed part of the source code of Roller and the time evaluation source code of TVM.
- I gathered and read some papers on neural networks and summarized the structures of them, so as to provide the benchmarks for our project.
- I added the functionality of resources allocation for the IOS and obtained about 10% improvement.

Picasso Lab, University of California, San Diego

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. The paper is in submission.
- I analyzed the problem of the embedding table placement on multi-GPU theoretically and proposed some possible algorithms for the load-balance memory allocation.
- I explored the existing approach to solving sharding problems such as using RL and proposed some potential ways to improve it.

Project: Combiantion 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.
- I analyzed the realistic data and got some insights about how the table placement would influence the inference time.
- I have completed some important components such as a prediction model to predict the time of table batched embedding.

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, I participated in the designing and revision of the course projects, conducted Q&A sessions in class, organized the quizzes, and worked on the final exam.

Skills

Programming Languages & Softwares: C, C++, Python, Linux

Python Packages: torch, scikit-learn

Languages: Mandarin (native), English (TOEFL iBT: 102/120), Korean (beginner)