Yeqi Huang

Email: yeqi.huang@ed.ac.uk

Personal Site: yeqi-huang.com

Education

- **University of Science and Technology of China** — Hefei, China
  Bachelor of Computer Science; GPA: 3.44/4.3 — *July 2017 - June 2021*
  Courses: *Operating Systems, Artificial Intelligence, Principle of Compiler, Computer Architecture, High Performance Architecture*
  Honors: *2017, 2018, 2019, 2020 Outstanding Student Scholarship*
- **University of Science and Technology of China** — Hefei, China
  Master of Computer Science; — *July 2021 - June 2022*
  Courses: *Computer Vision, Approximation algorithms, Distributed Algorithms, Parallel Programming*
- **University of Edinburgh** — Edinburgh, UK
  PhD of Computer Science; — *August 2023 - Present*
  Research Field: *AI-System, Distributed ML, ML-oriented architecture, Serverless*

Awards

International Awards:

- International Supercomputing Conference Student Cluster Competition Champion - 2021
- Asian Supercomputer Conference Student Cluster Competition First Prize - 2021

National Awards:

- Best Chinese Supercomputing Application of the Year - 2022
- Huawei Pioneer Developer (4 in China) - 2021
- National Compiler Designing Competition Second Prize - 2021

Projects

- **Dynamics of a tunable QED in quantum spin ice**:
  - Built tools to convert FORTRAN into modern C++ for further performance enhancement.
  - This work was presented at the APS conference and received guidance and recognition from Nobel Laureate Frank Wilczek.
- **ACM Gordon Bell Prize Nomination**:
  - This work was accepted at SC 2021 and received the Gordon Bell Prize nomination.
  - Participated in some optimization work based on the particle-in-cell method.
  - Provided visualization for this work.
- **AutoReader Project**:
  - Built a vector database for the latest ArXiv papers for daily semantic search and subscription.
  - Project link: https://autoreader.ed-aisys.com/.
- **More Projects Refer to Personal Website**: https://yeqi-huang.com

Recent Research

- **AgentWave: Multi-Agent Scheduling on Distributed System**: Seeking a chance on new hardware architecture.
- **VDBIndexBench: VectorDB Index benchmark in LLM**: Implemented different Index methods on GPU platform.
- **Few-Shot and Multi-Modal Model Training**: Conducting reinforcement learning research with Dartmouth and Harvard Medical School, focusing on advanced chart interpretation in biological literature. Employing CoT, ReFT, and MOE Pretrain techniques, the model surpasses GPT-4 by 23% in double-blind evaluations.
- **High-Performance KV Cache Assisted RAG System**: In-depth understanding and research on current RAG work, with some optimization ideas addressing slow retrieval speeds and long-text inference issues. Some of the tests have already been used in the AutoReader project.
- **Cerebras: Exploration of 2D-Mesh AI Chip**: Exploring research in architectural systems, I recently proposed an enhanced matrix multiplication algorithm surpassing Cannon. Implemented Transformer and Llama models with LLMoC, demonstrating 606× faster and 22× more energy-efficient GEMV, and **39× faster LLM decode with 1.7× better energy efficiency**. Submitted to OSDI 2025.

Publications

- **SC 21**: Symplectic structure-preserving particle-in-cell whole-volume simulation of tokamak plasmas to 111.3 trillion particles and 25.7 billion grids.
- **UKSys 2024**: InfiniTensor: A Tensor-Friendly, Efficient Parallel Programming Library for Accelerator-Centric Clusters
- **OSDI 2024**: ServerlessLLM: Locality-Enhanced Serverless Inference for Large Language Models

Skills

- **AI-Related**:
  - In-depth understanding of LLMs, with experience porting high-performance inference and training frameworks to various hardware platforms, including Apple Silicon, GPUs, and Cerebras.
  - Strong understanding of vector retrieval, having developed and tested graph and vector databases on multiple platforms.
  - In-depth knowledge of Multi-Agent applications and developed efficient development components for such applications.
- **Computer Science Related**:
  - Highly skilled in **C++, Python, Rust**; familiar with Go, JavaScript, and Latex
  - Proficient with **CUDA, Intel ONEAPI, OpenMP** and related parallel and distributed programming
  - Well-versed in **LLVM**, frequently participating in LLVM Forum online discussions
  - Extensive experience working with Linux, including usage of eBPF
  - Rich experience in distributed systems and distributed machine learning frameworks
  - Strong engineering development experience, mastering various compiler-related tools and static analysis tools
- **Physics & Math**:
  - Highly skilled in Computational Fluid Dynamics and Molecular Dynamics
  - Strong knowledge in numerical methods and linear algebra
  - Proficient in Quantum Mechanics and Quantum Electrodynamics, with some knowledge of quantum computing

Special Experience

- **Open Source Enthusiast**: Contributing to notable projects like GNOME, LAMMPS, and LLVM, I've used GitHub since 2019 to showcase my development journey and ideas.
- **Running a Science-Themed Cafe**: Leveraging software development income, I established Quantum Coffee near my school —a collaborative space encouraging students to explore and discuss scientific topics across disciplines.
- **UNICEF Charity Projects**: I have donated 10% of all personal income to children's charities since 2019.