Yeqi Huang

Email: yeqi.huang@ed.ac.uk

Personal Site: yeqi-huang.com

Education

- **University of Edinburgh** — Edinburgh, UK
  PhD in Computer Science — *August 2023 - Present*
  Research Field: *AI-Systems, Distributed ML, ML-oriented Architecture, Serverless*

- **University of Science and Technology of China** — Hefei, China
  Bachelor of Computer Science; The school of Gifted Young; — *July 2017 - June 2021*
  Courses: *Operating Systems, Artificial Intelligence, Principles of Compiler, Computer Architecture, High Performance Architecture*

Publications

1 Yao Fu, Leyang Xue, **Yeqi Huang**, et al. "ServerlessLLM: Locality-Enhanced Serverless Inference for Large Language Models." In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024.

2 Yao Fu, Yinsicheng Jiang, **Yeqi Huang**, et al. "MoE-CAP: Cost-Accuracy-Performance Benchmarking for Mixture-of-Experts Systems." *NeurIPS*, 2025.

3 Congjie He, **Yeqi Huang**, Pei Mu, et al. "WaferLLM: Large Language Model Inference at Wafer Scale." *Under Review*, OSDI 2025.

4 Xiao-Long Chen, Lin-Feng Wang, **Yeqi Huang**, et al. "Symplectic structure-preserving particle-in-cell whole-volume simulation of tokamak plasmas." In *SC21: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021.

Projects

- **QED Simulation on GPU**:
  - Implemented high-performance Monte-Carlo simulation achieving 1500x to 5000x speedup.
  - Received recognition from Nobel Laureate Frank Wilczek at APS conference.
  - Demonstrated expertise in CUDA programming and GPU micro-architecture optimization.
- **AutoReader Project**:
  - Built a vector database for the latest ArXiv papers for daily semantic search and subscription.(Full-stack AI application)
  - ***Implemented high-performance retrieval algorithm on GPU.***
  - Achieving ***60× faster indexing than NVIDIA cuVS*** library and ***13× faster retrieval than Milvus***.
  - Extended system to support bioRxiv and PubMed papers with specialized biology-focused features.
  - Biology patched version is demostrating in Dartmouth university, will submit to Nature this year.
- **Vector Search on Cerebras**:
  - Implemented KNN/ANN-based vector search algorithm on Cerebras chip.
  - Developing Retrieval based RAG and RetrievalAttention for long context LLM recently.
  - Achieved ***800× faster GEVV operations with 40× better energy efficiency***.
- **Spatial Accelerator Compiler (submitted to OSDI 2026)**:
  - Implement code generation tools on modern spatail accelerator such as Cerebras, Grok and Tenstorrent.
  - Designed Spatical IR that reduces communication overhead by 1000× compared to traditional collective APIs.
  - Achieved auto parallelism for LLM models, make clear decision about the placement of each operators.
- **SwarmPilot: AI Workload Scheduling System (submitted to OSDI 2026 Industry Track)**:
  - Designed and implemented an intelligent GPU cluster management system for AI workloads with multi-agent architecture.
  - Implemented two-stage adaptive profiling algorithm to automatically discover optimal system throughput sweet points.
  - Built complete ML pipeline from data collection, model training to online prediction for dynamic task scheduling.
  - Integrated with Ray and ComfyUI frameworks for distributed computing and workflow management.
- **BioVLM Research Project**:
  - Collaborated with Dartmouth and Harvard Medical School on biological literature interpretation.
  - Developed advanced Vision ability by adapting GRPO RL Training, CoT training, ReFT, and long context with RAG.
  - Achieved 23% improvement over GPT-4o in double-blind human evaluations.

- Achieved 10% improvement over GPT-5 in LabBench benchmark.

- **AI4Math - Sketchpad Project**:
  - Received grant from AI for Math Fund to develop Sketchpad system for formal mathematics (announcement)
  - Building system that automatically converts mathematical proofs into structured data using graph-based representation.
  - Enabling decomposition of proofs into individual statements for more precise auto-formalization.
- **More Projects (https://yeqi-huang.com/)**:
  - I have plenty of AI related projects on my personal page.

Talks

1 **Yeqi Huang**. "InfiniTensor: A Tensor-Friendly, Efficient Parallel Programming Library for Accelerator-Centric Clusters." *UKSys 2024*.

2 **Yeqi Huang**. "Why we need a new clustering benchmark in AI retrieval?" *International Workshop on Efficient Generative AI*, 2024.

Teaching Experience

- **Data and Visualization TA**:
  - Teaching assistant for Data and Visualization course, helping students learn large-scale data processing techniques.
  - Guided students through practical data analysis pipelines and visualization frameworks.
- **Machine Learning System TA**:
  - Teaching assistant for CUDA programming course, covering GPU architecture, CUDA, CuPy, and Triton.
  - Designed hands-on exercises and projects to help students master parallel programming concepts.

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

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.