Weimin Hu(胡伟民)

M.S. Candidate, Computer Science and Technology, Anhui University

Github: https://github.com/guoqingling

Email: wm.hu@siat.ac.cn

Research Interest

Scalable Scheduling for Large Language Model Serving
Analytical Performance Modeling of Multi-GPU Systems
AI/ML Systems Co-design with Memory&Interconnect

Education

2023.09 – 2026.06 (Expected)	M.S. in Computer Science, Anhui University
2024.03 – 2026.06 (Expected)	Joint Training, SIAT, CAS (Advisor: Prof. Zhibin Yu)
2017.09 – 2021.06	B.Eng. in IoT Engineering, Jiangxi Normal University

Publication

Weimin Hu, Zhibin Yu, et al. "An Analytical Model for Multi-viewer Caching Rendering Pipeline on Multi-GPU Systems." Manuscript in preparation for ASPLOS 2026.

Research Experience

MoE LLMServing Simulator (SIAT & Huawei Technologies) 2025.04 - present

– Problem: State-of-the-art LLM-serving simulators still use synchronous batching for the decode phase when the request rate is low; this inflates per-token latency (TPOT) because each user must wait for the slowest request in the batch.

– Method: Built an iteration-level, event-driven simulator that (i) replays real 8×H100 traces collected from a 671B-MoE production cluster and (ii) implements asynchronous expert scheduling with a dynamic memory pool to eliminate decode-phase synchronization.

– Result: Simulation shows that asynchronous expert scheduling cuts TPOT by 52 % and raises sustainable QPS by 1.9× under the same SLO, with trace-driven validation within ± 3 % of measured cluster performance.

Skills

C++17, Python, PyTorch, vLLM, Nsight Systems, Roofline model, GPU Simulator, LaTeX

Awards

Anhui University Outstanding Graduate Fellow (top 5 %), 2024
Anhui University First-class Scholarship (rank 1), 2024