Kaiwen Chen

https://kaiwenkevinn.github.io

+852 44281429 | kwchen@link.cuhk.edu.hk

Education

Aug 2024 - Present

The Chinese University of Hong Kong **Computer Science and Engineering（Doctorate）**

Supervisor: Prof. Hong Xu. Research Interests: Machine Learning Systems (MLSys, AI Infrastructure)

Sep 2020 - Jul 2024

Nanjing University **Software Engineering（Bachelor）**

GPA: 4.46/5.0, 3.82/4.0 (WES Algorithm). First-class Scholarship three times, Outstanding Student three times, Outstanding Graduate of Nanjing University

Research Interest

I am broadly interested in System Design for Machine Learning (AI Infra), including the following topics:

1. **Distributed Machine Learning Systems:** Efficient and scalable systems for LLM training, RL post-training, and inference, with an emphasis on resource management, KV-cache sharing, and performance simulation.
2. **Resource Management and Scheduling:** Designing resource management, workload scheduling, and scheduling algorithms for machine learning systems.

Internships

Jan 2026 - Aug 2026

Huawei Technologies **2012 Laboratories (University-enterprise cooperation)**

- Led the research and development of **Libra**, a resource management system for agentic RL post-training that addresses long-tailed rollouts and dynamic workload imbalance between rollout and training.
- Designed a global resource planner with elastic GPU reallocation and a causality-driven multi-level feedback queue (C-MLFQ) scheduler that routes trajectories using tool-return signals, achieving up to **4.2x higher throughput** and **2.7x faster reward convergence** on **48 NVIDIA A800 GPUs and 160 Ascend 910B NPU**.
- **Open-sourced** Libra (https://github.com/NetX-lab/Libra/tree/main, 200+ stars, 15+ forks)

Nov 2023 - May 2024

Microsoft Research Asia **Network Research Group**

- The scalability of BGP routing was studied, and the relevant contributions were included in the acknowledgments of the paper.
- Multi-layer LSTM was used to predict the growth of data center routing tables. Simulation tests of data center network equipment were designed and run to evaluate the behavior and performance when routing entries exceed limits. Improved route aggregation strategies and selective add-path schemes were proposed to alleviate routing table pressure, enhance stability, and improve resource utilization.

Publications

1. ***Kaiwen Chen***, Xin Tan, Jingzong Li, Hong Xu. “Libra: Efficient Resource Management for Agentic RL Post-Training”, in arXiv Preprint, June 2026.
2. ***Kaiwen Chen***, Xin Tan, Minchen Yu, Jingzong Li, Hong Xu. “ReasonCache: Accelerating Large Reasoning Model Serving through KV Cache Sharing”, **IEEE/ACM International Symposium on Quality of Service (IWQoS), 2026**
3. Yicheng Feng, Kin Hang Sew, Yuetao Chen, ***Kaiwen Chen***, Jingzong Li, Tianyuan Wu, Zheng Zhou, Peng Cheng, Chuan Wu, Wei Wang, Tsung-Yi Ho, Hong Xu. “Scalable and Efficient Simulation of LLM Training”, **International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2026.**

Projects

Resource Management for Agentic RL Post-Training

Preprint; first-author/lead work 2025.12 - 2026.08

- Analyzed the agentic RL workload and found that tool interactions produce highly variable trajectory lengths. Meanwhile, rollout and training have fundamentally different compute and memory requirements, causing any static GPU allocation between the two stages to become inefficient as the policy and workload evolve.
- Led the design and development of Libra. Libra periodically evaluates candidate system configurations and jointly reallocates GPUs between rollout and training through an elastic hybrid worker pool, allowing GPUs to switch roles without blocking ongoing execution. Within the rollout stage, its causality-driven scheduler uses runtime signals from tool results, such as payload size and execution failures, to identify trajectories likely to become long and migrate them to workers with more suitable parallel configurations. This approach avoids relying on inaccurate upfront length prediction.
- Evaluated Libra on three representative agentic RL workloads using 48 NVIDIA A800 GPUs and 160 Ascend 910B NPUs, achieving up to 4.2x higher throughput and 2.7x faster reward convergence than state-of-the-art baselines.

KV Cache Management for Large Reasoning Model

Published at IWQoS'26; *first-author/lead work* 2025.06 - 2025.12

- Analyzed large reasoning model traces and found that 15–40% of generated content exhibits redundant reasoning, producing similar KV-cache states that consume GPU memory and limit serving concurrency.
- Designed ReasonCache, an asynchronous coarse-to-fine filtering pipeline that identifies reusable KV-cache blocks and integrates reference-counted, zero-copy block sharing into PagedAttention-based serving systems.
- Improved serving throughput by 40–60% on average and up to 89.2%, while retaining over 98% of dense-attention accuracy and outperforming existing KV-cache management methods.

Simulator for Scalable Distributed Training

Published at SC'26*; core contributor* 2024.12 - 2025.06

- Identified three limitations of existing distributed-training simulators: dependence on full-cluster workload tracing, inaccurate collective-communication models, and failure to capture computation slowdowns caused by communication-computation overlap.
- Designed an ex-situ tracing method that reconstructs rank-specific execution graphs on a single GPU, complemented by a white-box NCCL communication model and an ML-based predictor for overlap-induced slowdown.
- Simulated GPT-175B training with 3D parallelism on 96 NVIDIA H800 GPUs in under 2 minutes, achieving an average step-time error of approximately 8%, about 3x lower than state-of-the-art simulators.

Awards and Honors

- **Full Postgraduate Scholarship**, CUHK 2024.08-2028.07
- **Outstanding Graduate**, Nanjing University 2024.05
- **First-class Scholarship,** Nanjing University 2024, 2023, 2022

Academic Services

Reviewer: AAAI 2026, AAAI 2027

Artifact Evaluation Committee: ATC 2026

Teaching

Teaching Assistant: CSCI 3150, Introduction to Operating Systems, CUHK. 2024 Fall
CSCI 3320, Fundamentals of Machine Learning, CUHK. 2025 Spring
CSCI 2040, Introduction to Python, CUHK. 2025 Fall
CSCI 3320, Fundamentals of Machine Learning, CUHK. 2026 Spring