陈凯文

https://kaiwenkevinn.github.io

+86 18810209048 | kwchen@link.cuhk.edu.hk

教育

2024.08 - 至今 **香港中文大学** **计算机科学与工程（博士）**

导师: Prof. Hong Xu. 研究方向: Machine Learning Systems (MLSys, AI Infrastructure)

2020.09 - 2024.07 **南京大学** **软件工程（本科）**

GPA: 4.46/5.0, 3.82/4.0 (WES Algorithm). 一等奖学金*3、优秀学生*3，南京大学优秀毕业生。

研究方向

我广泛关注**机器学习系统设计（AI Infra）**，具体包括以下方向：

1. 分布式机器学习系统：面向大语言模型（LLM）训练、强化学习后训练（RL post-training）及推理的高效可扩展系统，重点关注资源管理、KV缓存共享和性能模拟。
2. 资源管理与调度：为机器学习系统设计资源管理、工作负载调度及调度算法。

实习经历

2026.01 - 2026.08 **华为技术有限公司** **2012实验室 (校企合作)**

- 主导智能体强化学习后训练资源管理系统 Libra 的研究与开发，分析工具调用造成的轨迹长度波动、长尾任务以及训练过程中持续变化的跨阶段资源失衡问题。
- 设计全局资源规划器、弹性混合 GPU 资源池和基于工具返回信号的因果驱动调度器，使 GPU 能够在 rollout 与 training 阶段之间动态调整，并将长轨迹迁移至更适合的并行配置。在 **48 张 NVIDIA A800 GPU** 以及 **160 张 Ascend 910B NPU** 上完成系统评估，相比代表性基线最高提升 **4.2x 吞吐量**，并使训练达到同等奖励水平的速度最高提升 **2.7x**。
- **开源** Libra (https://github.com/NetX-lab/Libra/tree/main, 200+ stars, 15+ forks)

2023.11 - 2024.05 **微软亚洲研究院** **Network Research Group**

- 研究了BGP路由的可扩展性，利用多层LSTM预测了数据中心路由表的增长趋势。设计并实施了针对数据中心网络设备的仿真测试，以评估路由条目超出限制时的设备行为与性能。此外，提出了改进的路由聚合策略和选择性Add-Path方案，旨在缓解路由表压力、增强系统稳定性并提高资源利用率。

论文发表

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.**

项目经历

面向智能体强化学习后训练的资源管理

预印本；第一作者/主导工作 2025.12 – 2026.08

- 分析了智能体强化学习（Agentic RL）的工作负载，发现工具交互是导致轨迹长度高度可变的重要原因。同时，rollout和training两个阶段在计算与内存需求上存在根本性差异，导致两者之间任何静态的GPU分配方案都会随着策略和工作负载的演化而变得低效。
- 主导了Libra的设计与开发。Libra定期评估候选系统配置，并通过弹性混合工作线程池在rollout与training之间联合重分配GPU，使得GPU能够在不阻塞正在执行的任务的前提下切换角色。在rollout阶段内，其因果驱动调度器利用工具结果中的运行时信号（如负载大小和执行失败）来识别可能变长的轨迹，并将其迁移到具有更合适并行配置的工作线程上。该方法避免了对不准确的前置长度预测的依赖。在48块NVIDIA A800 GPU 和160块Ascend 910B NPU上，基于三种代表性的智能体RL工作负载对Libra进行了评估，相较于基线方案，**吞吐量最高提升4.2倍，奖励收敛速度最高提升2.7倍。**
- **开源Libra (https://github.com/NetX-lab/Libra/tree/main, 200+ stars, 15+ forks)**

面向大型推理模型的KV缓存管理

发表于 ***IWQoS'26（CCF-B）****；第一作者/主导工作* 2025.06 - 2025.12

- 分析了大型推理模型的运行轨迹，发现15%–40%的生成内容存在冗余推理，产生相似的KV缓存状态，从而消耗GPU内存并限制服务并发度。
- 设计了ReasonCache，一个异步的由粗到细的过滤流水线，用于识别可重用的KV缓存块，并将引用计数、零拷贝的块共享机制集成到基于PagedAttention的服务系统中。
- 平均提升服务吞吐量40%–60%，最高可达89.2%，同时保持超过98%的稠密注意力精度，并优于现有的KV缓存管理方法。

面向分布式训练的模拟器

发表于 ***SC'26（CCF-A）****；核心贡献者* 2024.12 - 2025.06

- 指出现有分布式训练模拟器的三个局限性：依赖全集群工作负载追踪、通信模型不准确，以及未能捕获通信-计算重叠所导致的计算变慢。
- 设计了一种外部追踪（ex-situ tracing）方法，可在单张GPU上重建每个rank的执行图，并辅以白盒NCCL通信模型和基于机器学习（ML）的重叠导致变慢预测器。
- 在96块NVIDIA H800 GPU上，对采用3D并行的GPT-175B训练进行了模拟，耗时不到2分钟，平均每步时间误差约为8%，相比最先进的模拟器降低了约3倍。

荣誉与奖项

- **博士研究生奖学金**, 香港中文大学 2024.08 - 2028.07
- **优秀毕业生**, 南京大学 2024.05
- **人民奖学金一等奖,** 南京大学 2022, 2023, 2024

专业服务

审稿人: AAAI 2026, AAAI 2027

Artifact Evaluation Committee: ATC 2026

技术能力

深度学习: PyTorch, Verl, vLLM, SGLang, Megatron, CUDA, NCCL

编程语言: Python, C++, Java, HTML, CSS, JavaScript, LaTex

英语: IELTS 7.5