# Weihao **Cui**, 崔炜皞

weihao-cui.com github.com/Raphael-Hao linkedin.com/in/weihao-cui-b5a625a5
raphaelhao@outlook.com weihao@sjtu.edu.cn
No.800 Dongchuan Road, Minhang District, Shanghai, China
Apr. 1992, Nantong, Jiangsu, China

**Bio.** I am a postdoctoral research fellow working with Prof. Bingsheng He and Prof. Minyi Guo. I am jointly supported by the National University of Singapore and Shanghai Jiao Tong University and work closely with Prof. Quan Chen and Dr. Han Zhao. I obtained my Ph.D. degree at the Department of Computer Science and Engineering (CSE), Shanghai Jiao Tong University (SJTU), China, working under the guidance of Prof. Quan Chen and Prof. Minyi Guo on **AI Infrastructure and Cloud Computing Systems**.

To date, I have published **21** CCF-A−tier conference and journal papers, including **8** as first author. These works have achieved both academic impact–earning distinctions such as the **MLSys 2025 Outstanding Paper Honorable Mention and SC 2021 Best Reproducibility Advancement Award Finalist**–and industrial impact, with deployments at ByteDance, Ant Group, and other organizations. In recognition of my sustained contributions to machine learning systems, I was selected for the **CCF Doctoral Dissertation Incentive Program (2025)**.

**Research Interests.**
My research spans several topics at the intersection of AI infrastructures and computing systems :

- **AI Infrastructure** : Designed and optimized systems for efficient training and serving of diverse AI workloads, including large language models, embodied agents, and multimodal architectures.
- **Cloud Computing Systems** : Developed platforms for GPU serverless computing, cloud gaming, and accelerated datacenter operations, with a focus on performance and scalability.

My current work focuses on optimizing AI systems, with a particular emphasis on large language models (LLMs). To the best of my knowledge, I was the first to propose improving the inference efficiency of Mixture-of-Experts (MoE) models by leveraging expert placement strategies for load balancing. My ongoing project, *Yoda*, introduces a new paradigm for serving LLMs with Service Level Objective (SLO) guarantees. It advances beyond existing techniques such as disaggregated serving and chunked prefill, and **has been integrated into mainstream LLM serving frameworks like SGLang** BLOG Link .

## Career

| | |
|---|---|
| June 2025 – Now | **Visiting Scholar** in *Ant Group*<br>Collaborating on Large-Scale LLM Training |
| Oct 2023 – Now | **Postdoctoral Researcher** in *National University of Singapore* and *Shanghai Jiao Tong University*<br>Xtra Computing Group and Emerging Parallel Computing Center (EPCC)<br>*Advisor : Prof. Minyi Guoand Prof. Bingsheng He* |
| Dec 2021 – Jan 2023 | **Research Intern** in *Microsoft Research Asia*<br>Systems and Engineering Group in Shanghai<br>*Advisor : Zhenhua Han* |

## Education

| | |
|---|---|
| Sep 2020 – Sep 2023 | **Ph.D.** in *Shanghai Jiao Tong University*, Shanghai<br>Emerging Parallel Computing Center (EPCC)<br>*Advisor : Prof. Quan Chen* |
| Sep 2017 – Apr 2020 | **Master** in *Shanghai Jiao Tong University*, Shanghai<br>Emerging Parallel Computing Center (EPCC)<br>*Advisor : Prof. Quan Chen*<br>GPA : 3.73/4.0 |
| Sep 2011 – Jun 2015 | **Bachelor of Engineering** in *Shanghai Jiao Tong University*, Shanghai<br>Majors in Ocean Engineering |

## Projects & Funds

| | |
|---|---|
| October 2024-<br>September 2027 | A Unified Co-location Technique for Training and Inference of Large Language Models, SJTU, 200KRMB<br>Principal Investigator, responsible for the whole project<br>Natural Science Foundation of Shanghai Municipality (24ZR1430500) |
| December 2024-<br>November 2027 | A System for Real-Time Resource Isolation and Efficient Scheduling on Heterogeneous Computing Platforms, SJTU & ICT & XJTU & Cecloud & USTC & SZU & Lenovo & China Telecom, 24MRMB<br>Project Contact Person, responsible for sub project 3 : Hybrid Task Scheduling for Diverse QoS Guarantees<br>National Key Research and Development Program of China (2024YFB4505700) |

## Honors & Awards

| | |
|---|---|
| **Dec** 2025 | CCF Doctoral Dissertation Incentive Program **(Top 10 Nationwide)**. |
| **Aug** 2025 | APPT CCF Distinguished Paper Award **(2/74, Worldwide)**. |
| **May** 2025 | MLSys Outstanding Paper Honorable Mention **(5/271, Worldwide; only 1 from China)**. |
| **Sep** 2024 | CCF Doctoral Dissertation Incentive Program on HPC **(Top 5 Nationwide)**. |
| **Sep** 2023 | Siyuan Overseas Joint Postdoctoral Training Program of Shanghai Jiao Tong University. **(On par with the National Postdoctoral Program for Innovative Talent, 360K RMB per Year)**. |
| **Dec** 2023 | ICPADS Best Paper Award **(1/196, Worldwide)**. |
| **Feb** 2023 | Microsoft Stars of Tomorrow. |
| **Sep** 2021 | SC 2021 Best Reproducibility Advancement Finalists **(3/365, Worldwide; only 1 from China)**. |
| **Sep** 2021 | National Fellowship of Shanghai Jiao Tong University. |
| **May** 2020 | Tencent Rhino-bird Elite Training Program. |

## Publications

* denotes equal contribution; † indicates the corresponding author.

**Preprint.**

- **Weihao Cui**, Ziyi Xu, Han Zhao, Quan Chen, Zijun Li, Bingsheng He, Minyi Guo. "Efficient Function-as-a-Service for Large Language Models with TIDAL." **Submitted to OSDI 2026, CCF-A**.
- Ao Xu, Han Zhao, **Weihao Cui**, Quan Chen, Yukang Chen, Shulai Zhang, Shuang Chen, Jiemin Jiang, Zhibin Yu, and Minyi Guo. "Harli : SLO-Aware Co-location of LLM Inference and PEFT-based Finetuning on Model-as-a-Service Platforms." **Submitted to ISCA 2026, CCF-A**.
- Shulai Zhang, Ao Xu, Quan Chen, Han Zhao, **Weihao Cui**, Ningxin Zheng, Haibin Lin, Xin Liu, Minyi Guo. "Boosting Embodied AI Agents through Perception-Generation Disaggregation and Asynchronous Pipeline Execution." **Submitted to EuroSys 2026, CCF-A**.
- Shulai Zhang, Ao Xu, Quan Chen, Han Zhao, Weihao Cui, Ningxin Zheng, Haibin Lin, Xin Liu, and Minyi Guo. "Boosting Embodied AI Agents through Perception-Generation Disaggregation and Asynchronous Pipeline Execution" **arXiv :2509.09560, 2025**.

**Conference.**

- Yushi Liu, Yikang Ruan, Letian Ruan, Zijun Li, Sen Gao, **Weihao Cui**, Quan Chen, Shuo Quan, Jie Wu, Bingsheng He, Minyi Guo. "FaaSBoard : Efficient Graph Processing with a Disaggregated Architecture on Serverless Services." **SIGMOD 2026 Major Revision, CCF-A**.
- Chunyu Xue, **Weihao Cui**, Quan Chen, Chen Chen, Han Zhao, Shulai Zhang, Linmei Wang, Yan Li, Limin Xiao, Weifeng Zhang, Jing Yang, Bingsheng He, Minyi Guo. "Arena : Efficiently Training Large Models via Dynamic Scheduling and Adaptive Parallelism Co-Design." **EuroSys 2026 Major Revision, CCF-A**.
- **Weihao Cui**, Yukang Chen, Han Zhao, Ziyi Xu, Quan Chen, Xusheng Chen, Yangjie Zhou, Shixuan Sun, Minyi Guo"Optimizing SLO-oriented LLM Serving with PD-Multiplexing." **ASPLOS 2026 Major Revision, CCF-A**.
- Chunyu Xue, Yi Pan, **Weihao Cui**, Quan Chen, Shulai Zhang, Bingsheng He, Minyi Guo. "MuxTune : Efficient Multi-Task LLM Fine-Tuning in Multi-Tenant Datacenters via Spatial-Temporal Backbone Multiplexing." **NSDI 2026, CCF-A**.
- **Weihao Cui** Ji Zhang, Han Zhao, Chao Liu, Wenhao Zhang, Jian Sha, Quan Chen, Bingsheng He, Minyi Guo. "Flare : Anomaly Diagnostics for Divergent LLM Training in GPU Clusters of Thousand-Plus Scale." **NSDI 2026, CCF-A**.
- Han Zhao*, **Weihao Cui***, Zeshen Zhang, Wenhao Zhang, Jiangtong Li, Quan Chen, Pu Pang, Zijun Li, Zhenhua Han, Yuqing Yang, Minyi Guo."LEGO : Supporting LLM-enhanced Games with One Gaming GPU." **HPCA 2026, CCF-A**.
- Yangjie Zhou, Honglin Zhu, Qian Qiu, **Weihao Cui**, Zihan Liu, Cong Guo, Siyuan Feng, Jintao Meng, Haidong Lan, Jingwen Leng, Wenxi Zhu, Minwen Deng. "A Sample-Free Compilation Framework for Efficient Dynamic Tensor Computation." **SC 2025, CCF-A**.
- Yangjie Zhou, Wenting Shen, Jingwen Leng, Shuwen Lu, Zihan Liu, **Weihao Cui**, Zhendong Zhang, Wencong Xiao, Baole Ai, Wei Lin, Deze Zeng, Yun Liang, Quan Chen, Ning Liu, Minyi Guo. "Voyager : Input-Adaptive Algebraic Transformations for High-Performance Graph Neural Networks." **ASPLOS 2025, CCF-A**.
- Han Zhao, Yiying Xiang, Yu Liu, Xiaochun Ye, Deze Zeng, Jing Yang, **Weihao Cui**, Quan Chen, Jingwen Leng, Minyi Guo. "DACO : Unlocking Latent Dataflow Opportunities in Edge-Side SIMT Accelerators." **APPT 2025, CCF Distinguished Paper Award**.
- Shulai Zhang, Ningxin Zheng, Haibin Lin, Ziheng Jiang, Wenlei Bao, Chengquan Jiang, Qi Hou, **Weihao Cui**, Size Zheng, Li-Wen Chang, Quan Chen, Xin Liu. "Comet : Fine-grained Computation-communication Overlapping for Mixture-of-Experts." **MLSys 2025, Outstanding Paper Honorable Mention**.
- Shulai Zhang, Quan Chen, **Weihao Cui**, Han Zhao, Chunyu Xue, Zhen Zheng, Wei Lin, Minyi GUo. "Improving GPU Sharing Performance through Adaptive Bubbleless Spatial-Temporal Sharing." **Eurosys 2025, CCF-A**.
- Zihan Liu, Xinhao Luo, Junxian Guo, Wentao Ni, Yangjie Zhou, Yue Guan, Cong Guo, **Weihao Cui**, Yu Feng, Minyi Guo, Yuhao

Zhu, Minjia Zhang, Chen Jin, Jingwen Leng. "VQ-LLM : High-performance Code Generation for Vector Quantization Augmented LLM Inference." **HPCA 2025, CCF-A**.

- Jiagan Cheng, Yilong Zhao, Zijun Li, Quan Chen, **Weihao Cui**, Minyi Guo. "Microless : Cost-Efficient Hybrid Deployment of Microservices on IaaS VMs and Serverless." **ICPADS 2023, Best Paper, CCF-C**.
- Binghao Chen, Han Zhao, **Weihao Cui**, Yifu He, Shulai Zhang, Quan Chen, Zijun Li, Minyi Guo. "Maximizing the Utilization of GPUs Used by Cloud Gaming through Adaptive Co-location with Combo." **SoCC 2023, CCF-B**.
- Yangjie Zhou, Yaoxu Song, Jingwen Leng, Zihan Liu, **Weihao Cui**, Zhendong Zhang, Cong Guo, Quan Chen, Li Li, Minyi Guo. "AdaptGear : Accelerating GNN Training via Adaptive Subgraph-Level Kernels on GPUs." **CF 2023, CCF-C**.
- **Weihao Cui**, Zhenhua Han, Lingji Ouyang, Yichuan Wang, Ningxin Zheng, Lingxiao Ma, Yuqing Yang, Fan Yang, Jilong Xue, Lili Qiu, Lidong Zhou, Quan Chen, Haisheng Tan, Minyi Guo. "Optimizing Dynamic Neural Networks with Brainstorm" **OSDI 2023, CCF-A**.
- Shulai Zhang, **Weihao Cui**, Quan Chen, Zhengnian Zhang, Yue Guan, Jingwen Leng, Chao Li, Minyi Guo. "PAME : Precision-Aware Multi-Exit DNN Serving for Reducing Latencies of Batched Inferences." **ICS 2022, CCF-B**.
- **Weihao Cui**, Han Zhao, Quan Chen, Hao Wei, Zirui Li, Deze Zeng, Chao Li, Minyi Guo. "DVABatch : Diversity-aware Multi-Entry Multi-Exit Batching for Efficient Processing of DNN Services on GPUs." **ATC 2022, CCF-A**.
- Han Zhao, **Weihao Cui**, Quan Chen, Youtao Zhang, Yanchao Lu, Chao Li, Jingwen Leng, Minyi Guo. "Tacker : Tensor-CUDA Core Kernel Fusion for Improving the GPU Utilization while Ensuring QoS." **HPCA 2021, CCF-A**.
- **Weihao Cui**, Han Zhao, Quan Chen, Ningxin Zheng, Jingwen Leng, Jieru Zhao, Zhuo Song, Tao Ma, Yong Yang, Chao Li, and Minyi Guo. "Enable Simultaneous DNN Services Based on Deterministic Operator Overlap and Precise Latency Prediction." **SC 2021, Best Reproducibility Advancement Award Finalists, CCF-A**.
- Han Zhao, **Weihao Cui**, Quan Chen, Jieru Zhao, Jingwen Leng, and Minyi Guo. "Exploiting Intra-SM Parallelism in GPUs via Persistent and Elastic Blocks." **ICCD 2021, CCF-B**.
- Han Zhao, **Weihao Cui**, Quan Chen, Jingwen Leng, Kai Yu, Deze Zeng, Chao Li, and Minyi Guo. "CODA : Improving Resource Utilization by Slimming and Co-locating DNN and CPU Jobs." **ICDCS 2020, CCF-B**.
- **Weihao Cui**, Mengze Wei, Quan Chen, Xiaoxin Tang, Jingwen Leng, Li Li, and Mingyi Guo. "Ebird : Elastic Batch for Improving Responsiveness and Throughput of Deep Learning Services." **ICCD 2019, CCF-B**.
- Wei Zhang, **Weihao Cui**, Kaihua Fu, Quan Chen, Daniel Edward Mawhirter, Bo Wu, Chao Li, and Minyi Guo. "Laius : Towards Latency Awareness and Improved Utilization of Spatial Multitasking Accelerators in Datacenters." **ICS 2019, CCF-B**.

**Journal.**

- Yifei Liu, Chen Chen, Qiang Wang, Yu Feng, **Weihao Cui**, Quan Chen, Minyi Guo. "Ares : Fair and Efficient Scheduling of Deep Learning Jobs with Elastic Fair Queuing." **TACO 2025, CCF-A**.
- Pengyu Yang*, **Weihao Cui***, Chunyu Xue, Han Zhao, Chen Chen, Quan Chen, Jing Yang, Minyi Guo"Taming Flexible Job Packing in Deep Learning Training Clusters." **TACO 2025, CCF-A**.
- Han Zhao*, **Weihao Cui***, Quan Chen, Shulai Zhang, Zijun Li, Jingwen Leng, Chao Li, Deze Zeng, Minyi Guo. "Towards Fast Setup and High Throughput of GPU Serverless Computing." **TACO 2025, CCF-A**.
- Yifu He, Han Zhao, **Weihao Cui**, Shulai Zhang, Quan Chen, Minyi Guo. "ARACHNE : Optimizing Distributed Parallel Applications with Reduced Inter-Process Communication." **TACO 2025, CCF-A**.
- Han Zhao, Junxiao Deng, **Weihao Cui**, Quan Chen, Youtao Zhang, Deze Zeng, Minyi Guo. "Adaptive Kernel Fusion for Improving the GPU Utilization while Ensuring QoS." **TC 2024, CCF-A**.
- Cong Guo, Fengchen Xue, Jingwen Leng, Yuxian Qiu, Yue Guan, **Weihao Cui**, Quan Chen, Minyi Guo. "Accelerating Sparse DNNs based on Tiled Gemm." **TC 2024, CCF-A**.
- Han Zhao, **Weihao Cui**, Quan Chen, Jingwen Leng, Deze Zeng, Minyi Guo. "Improving Cluster Utilization through Adaptive Resource Management for DNN and CPU Jobs Co-location" **TC 2023, CCF-A**.
- Han Zhao, **Weihao Cui**, Quan Chen, Minyi Guo. "ISPA : Exploiting Intra-SM Parallelism in GPUs via Fine-grained Resource Management." **TC 2022, CCF-A**.
- **Weihao Cui**, Quan Chen, Han Zhao, Mengze Wei, Xiaoxin Tang, and Minyi Guo. "$E^2$bird : Enhanced Elastic Batch for Improving Responsiveness and Throughput of Deep Learning Services." **TPDS 202, CCF-A**.
- Wei Zhang, Quan Chen, Ninxing Zheng, **Weihao Cui**, Kaihua Fu, and Minyi Guo. "Towards QoS-awareness and Improved Utilization of Spatial Multitasking GPUs." **TC 2021, CCF-A**.

## Teaching

| | | | |
|---|---|---|---|
| 2022 | **AI Computing System** | Shanghai Jiao Tong University | Role : Teaching Assistant |
| 2021 | **Parallel Computing** | Shanghai Jiao Tong University | Role : Teaching Assistant |
| 2020 | **Parallel Computing** | Shanghai Jiao Tong University | Role : Teaching Assistant |

## Professional Services

| | |
|---|---|
| **2026** | ICPP Web Chair |
| **2025** | ChinaSys TPC member |
| **2025** | ICCD TPC member |
| **2026** | Eurosys Shadow TPC |
| **2025** | JCST reviewer |
| **2024** | ATC Artifact Evaluation PC |
| **2024** | OSDI Artifact Evaluation PC |
| **2024** | PMAM TPC member |
| **2024** | Bigcom TPC member |
| **2023** | ATC Artifact Evaluation PC |
| **2023** | OSDI Artifact Evaluation PC |
| **2023** | ICA3PP TPC member |
| **2022** | ATC Artifact Evaluation PC |
| **2022** | OSDI Artifact Evaluation PC |
| **2020** | FCS reviewer |

(last update : 5 Jan. 2026)