Qinghao Hu

S-Lab for Advanced Intelligence
Nanyang Technological University
ABN-02b-11, 61 Nanyang Avenue, Singapore 637335

E-Mail: qinghao.hu@ntu.edu.sg
Tel.: (+65) 8305 3277
Homepage: https://tonyhao.xyz

EMPLOYMENT

Research Assistant Professor *Jan. 2024 ~ Now*
S-Lab, *NTU, Singapore*

Academic Guest *Feb. 2024 ~ Apr. 2024*
Systems Group, *ETH Zürich, Switzerland*

EDUCATION

Nanyang Technological University, Singapore *2020 ~ 2023*
Ph.D. in Computer Science
Supervisor: Prof. Tianwei Zhang and Prof. Yonggang Wen

National University of Singapore, Singapore *2018 ~ 2020*
Master in Electrical Engineering

Zhejiang University, China *2014 ~ 2018*
Bachelor in Electrical Engineering

RESEARCH INTEREST

- Systems for Large Models
- Datacenter Management and Scheduling
- Machine Learning for Systems

PUBLICATION

Conference & Journal Papers

1. **Characterization of Large Language Model Development in the Datacenter**
Qinghao Hu*, Zhisheng Ye*, Zerui Wang*, Guoteng Wang, Meng Zhang, Qiaoling Chen, Peng Sun, et al.
[NSDI '24] *USENIX Symposium on Networked Systems Design and Implementation*

2. **Hydro: Surrogate-Based Hyperparameter Tuning Service in Datacenters**
Qinghao Hu, Zhisheng Ye, Meng Zhang, Qiaoling Chen, Peng Sun, Yonggang Wen, Tianwei Zhang
[OSDI '23] *USENIX Symposium on Operating Systems Design and Implementation*

3. **Lucid: A Non-Intrusive, Scalable and Interpretable Scheduler for Deep Learning Training Jobs**
Qinghao Hu*, Meng Zhang*, Peng Sun, Yonggang Wen, Tianwei Zhang
[ASPLOS '23] *Architectural Support for Programming Languages and Operating Systems*
Distinguished Paper Award

4. **Primo: Practical Learning-Augmented Systems with Interpretable Models**
Qinghao Hu, Harsha Nori, Peng Sun, Yonggang Wen, Tianwei Zhang
[ATC '22] *USENIX Annual Technical Conference*

5. **Characterization and Prediction of Deep Learning Workloads in Large-Scale GPU Datacenters**
Qinghao Hu, Peng Sun, Shengen Yan, Yonggang Wen, Tianwei Zhang
[SC '21] *International Conference for High Performance Computing, Networking, Storage, and Analysis*

6. **Deep Learning Workload Scheduling in GPU Datacenters: A Survey**
Zhisheng Ye*, Wei Gao*, Qinghao Hu*, Peng Sun, Xiaolin Wang, Yingwei Luo, Tianwei Zhang, et al.
[CSUR '24] *ACM Computing Surveys*

7. **TorchGT: A Holistic System for Large-scale Graph Transformer Training**
Meng Zhang*, Jie Sun*, Qinghao Hu, Peng Sun, Zeke Wang, Yonggang Wen, Tianwei Zhang
[SC '24] *International Conference for High Performance Computing, Networking, Storage, and Analysis*

8. **Sylvie: 3D-adaptive and Universal System for Large-scale Graph Neural Network Training**
Meng Zhang, Qinghao Hu, Cheng Wan, Haozhao Wang, Peng Sun, Yonggang Wen, Tianwei Zhang
[ICDE '24] *IEEE International Conference on Data Engineering*

9. **FedDSE: Distribution-aware Sub-model Extraction for Federated Learning over Resource-constrained Devices**
Haozhao Wang, Yabo Jia, Meng Zhang, Qinghao Hu, Hao Ren, Peng Sun, Yonggang Wen, Tianwei Zhang
[WWW '24] *The Web Conference*

Under Review

1. **DeltaServe: Multi-Tenant Language Model Serving via Delta Compression**
Xiaozhe Yao, Qinghao Hu, Ana Klimovic
[Preprint] *Submitted to a Conference*

2. **LoongTrain: Efficient Training of Long-Sequence LLMs with Head-Context Parallelism**
Diandian Gu, Peng Sun, Qinghao Hu, Ting Huang, Xun Chen, Yingtong Xiong, Guoteng Wang, Qiaoling Chen, Shangchun Zhao, Jiarui Fang, Yonggang Wen, Tianwei Zhang, Xin Jin, Xuanzhe Liu
[Preprint] *Submitted to a Conference*

3. **InternEvo: Efficient Long-Sequence Large Language Model Training via Hybrid Parallelism and Redundant Sharding**
Qiaoling Chen, Diandian Gu, Guoteng Wang, Xun Chen, Yingtong Xiong, Ting Huang, Qinghao Hu, Xin Jin, Yonggang Wen, Tianwei Zhang, Peng Sun
[Preprint] *Submitted to a Conference*

4. **AMSP: Super-Scaling LLM Training via Advanced Model States Partitioning**
Qiaoling Chen, Qinghao Hu, Zhisheng Ye, Guoteng Wang, Peng Sun, Yonggang Wen, Tianwei Zhang
[Preprint] *Submitted to a Conference*

AWARD

ML and Systems Rising Stars	*2024*
Outstanding Ph.D. Thesis Award	*2024*
National Scholarship for Outstanding International Graduates	*2024*
Google Ph.D. Fellowship	*2023*
Distinguished Paper Award of ASPLOS '23	*2023*
Youth Outstanding Paper Award of WAIC '23	*2023*
Student Travel Grant of OSDI '23	*2023*
Best Undergraduate Thesis Award	*2018*
Outstanding Graduates of Zhejiang University	*2018*

PROFESSIONAL SERVICE

[HASP '24] HASP Workshop (co-located with MICRO '24)	Publicity Chair
[EuroSys '24] EuroSys Conference	Shadow Committee Member
[EuroSys '23] EuroSys Conference	Shadow Committee Member
[OSDI '22] USENIX Symposium on Operating Systems Design and Implementation	AE Committee Member
[ATC '22] USENIX Annual Technical Conference	AE Committee Member

[**EuroSys '22**] EuroSys Conference — AE Committee Member

[**SOSP '21**] ACM Symposium on Operating Systems Principles — AE Committee Member

TALK

Characterization of Large Language Model Development in the Datacenter

Huawei, Shanghai China — *Jun. 2024*

NSDI, Santa Clara United States — *Apr. 2024*

Hydro: Surrogate-Based Hyperparameter Tuning Service in Datacenters

ChinaSys, Wuhan China — *Jul. 2023*

OSDI, Boston United States — *Jul. 2023*

Lucid: A Non-Intrusive, Scalable and Interpretable Scheduling System

Huawei, Beijing China — *May. 2023*

MLSys Seminar Singapore — *Apr. 2023*

ASPLOS, Vancouver Canada — *Mar. 2023*

Primo: Practical Learning Systems with Interpretable Models

ChinaSys, Nanjing China — *Dec. 2022*

ATC, Carlsbad California United States — *Jul. 2022*

Scheduling in Large-Scale GPU Datacenters

National University of Singapore — *Jan. 2022*

Characterization and Prediction of DL Workloads in Datacenters

SC, St. Louis Missouri United States — *Nov. 2021*

Cluster Scheduling for Deep Learning

S-Lab for Advanced Intelligence, Singapore — *Apr. 2021*