Qinghao Hu

HAN Lab
Massachusetts Institute of Technology
50 Vassar Street, Cambridge, MA 02139

E-Mail: qinghao@mit.edu
Tel.: (857)209-1101
Homepage: https://tonyhao.xyz

EXPERIENCE

Massachusetts Institute of Technology, United States *Aug. 2024 ~ Present*
Postdoctoral Associate, supervised by Dr. Song Han

ETH Zürich, Switzerland *Feb. 2024 ~ Apr. 2024*
Academic Visitor, supervised by Dr. Ana Klimovic

Nanyang Technological University, Singapore *Jan. 2024 ~ Aug. 2024*
Research Assistant Professor, supervised by Dr. Tianwei Zhang

EDUCATION

Nanyang Technological University, Singapore *2020 ~ 2023*
Ph.D. in Computer Science, supervised by Dr. Tianwei Zhang and Dr. 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
- Algorithm-System Co-design

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*
Best Undergraduate Thesis Award	*2018*
Outstanding Graduates of Zhejiang University	*2018*

PUBLICATION

Conference & Journal Papers

1. **Jet-Nemotron: Efficient Language Model with Post Neural Architecture Search**
Yuxian Gu, Qinghao Hu, Haocheng Xi, Junyu Chen, Shang Yang, Song Han, Han Cai
[NeurIPS '25] *Conference on Neural Information Processing Systems*

2. **Scaling up Reasoning to Long Videos in VLMs**
Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, Song Han
[NeurIPS '25] *Conference on Neural Information Processing Systems*

3. **Zeppelin: Balancing Variable-length Workloads in Data Parallel Large Model Training**
Chang Chen, Tiancheng Chen, Jiangfei Duan, Qianchao Zhu, Zerui Wang, Qinghao Hu, Peng Sun, Xiuhong Li, Chao Yang, Torsten Hoefler
[EuroSys '26] *EuroSys Conference*

4. **Sailor: Automating Distributed Training over Dynamic, Heterogeneous, and Geo-distributed Clusters**
Foteini Strati, Zhendong Zhang, George Manos, Ixeia Sánchez Périz, Qinghao Hu, Tiancheng Chen, Berk Buzcu, Song Han, Pamela Delgado, Ana Klimovic
[SOSP '25] *ACM Symposium on Operating Systems Principles*

5. **LServe: Efficient Long-sequence LLM Serving with Unified Sparse Attention**
Shang Yang*, Junxian Guo*, Haotian Tang, Qinghao Hu, Guangxuan Xiao, Jiaming Tang, Yujun Lin, Zhijian Liu, Yao Lu, Song Han
[MLSys '25] *Annual Conference on Machine Learning and Systems*

6. **LongVILA: Scaling Long-Context Visual Language Models for Long Videos**
Yukang Chen*, Fuzhao Xue*, Dacheng Li*, Qinghao Hu*, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, Song Han
[ICLR '25] *International Conference on Learning Representations*

7. **DeltaServe: Multi-Tenant Language Model Serving via Delta Compression**
Xiaozhe Yao, Qinghao Hu, Ana Klimovic
[EuroSys '25] *EuroSys Conference*

8. **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*

9. **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*

10. **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

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

12. **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*

13. **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*

14. **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*

15. **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*

16. **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. **Taming the Long-Tail: Efficient Reasoning RL Training with Adaptive Drafter**
Qinghao Hu*, Shang Yang*, Junxian Guo, Xiaozhe Yao, Chuang Gan, Ana Klimovic, Song Han
[Preprint] *Submitted to a Conference*

2. **Enhancing Deep Learning Schedulers with Large Language Models**
Zerui Wang*, Qinghao Hu*, Ana Klimovic, Tianwei Zhang, Yonggang Wen, Dahua Lin, Peng Sun
[Preprint] *Submitted to a Conference*

3. **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*

4. **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*

5. **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*

PROFESSIONAL SERVICE

[ICLR '26] International Conference on Learning Representations	Committee Member
[CVPR '25-ELVM] Efficient Large Vision Models Workshop	Organizer
[ICLR '25] International Conference on Learning Representations	Committee Member
[EuroSys '25] EuroSys Conference	Shadow Committee Member
[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*