Haoran Qiu

PhD Candidate
Computer Science, University of Illinois Urbana-Champaign

Contact Information

240 Coordinated Science Lab
1308 W Main St., Urbana, IL 61801

Website: https://haoran-qiu.com
E-mail: haoranq4@illinois.edu

Research Interests

Distributed Systems, Deep Learning Systems, Cloud Computing, ML for Systems

Education

University of Illinois Urbana-Champaign

Ph.D. Candidate, Computer Science, May 2024 (expected)

- Thesis: Cloud Datacenter Management with Efficient and Robust Online Learning
- Advisor: Prof. Ravishankar K. Iyer
- Thesis Committee: Tamer Başar, Klara Nahrstedt, Indranil Gupta, Onur Mutlu

The University of Hong Kong

B.Eng. in Computer Science with *First Class Honours*, May 2019

- Thesis: Building Fast and Robust Blockchain Systems for AI Applications
- Advisor: Prof. Heming Cui

Honors and Awards

- ML and Systems Rising Stars, MLCommons, 2023
- Mavis Future Faculty Fellowship, UIUC, 2023–24
- UIUC CS PhD Fellowship, UIUC, 2023–24
- Best Presentation Award, Workshop on Composable Systems (IPDPS), 2022
- Yunni & Maxine Pao Memorial Fellowship, UIUC, 2021
- Conference Presentation Award, UIUC, 2020
- Travel Grants: MLSys 2024, USENIX OSDI/ATC 2023, DSN 2022, ACM SIGMETRICS 2021
- Best Undergraduate Thesis 2nd Runner-up, HKU, 2019
- Dean's Honour List, HKU, 2016–19
- International Student Academic Excellence Award, University of Wisconsin-Madison, 2018
- Lee Shau Kee Scholarships for Student Enrichment, HKU, 2017
- Honorable Mention in Mathematical Contest In Modeling, COMAP, 2017
- HKU Foundation Scholarships For Outstanding Students, 2015–19

Professional Experience

PhD Software Engineer Intern, **Google SRG**, Mountain View, CA. May 2023 – Aug. 2023
Managers: Dr. Phitchaya Mangpo Phothilimthana (DeepMind), Dr. Yawen Wang
Designed and implemented power-aware datacenter capacity management with coordinated inter-cluster scheduling and intra-cluster bin-packing.

Visiting Researcher, **IBM Research**, Yorktown Heights, NY. Sept. 2022 – Dec. 2022
Managers: Dr. Chen Wang, Dr. Hubertus Franke
*Designed and implemented AWARE (published at **ATC 2023** [C10]), an extensible framework that provides robust and reliable deployment of online learning agents in production cloud platforms.*

PhD Software Engineer Intern, **Google**, Sunnyvale, CA. May 2022 – Aug. 2022
Managers: Dr. Qiaobin Fu, Dr. Joon Kang
*Designed and implemented network-aware scheduling of best-effort jobs for Borg cluster manager by extending FIRM (published at **OSDI 2020** [C8]).*

Research Intern, **Microsoft Research**, Redmond, WA (Virtual). May 2021 – Aug. 2021
Managers: Dr. Pulkit Misra, Dr. Íñigo Goiri
*Developed an analytical model to predict performance improvement for workloads with overclocking and an overclocking and power management framework (published at **ISCA 2024** [C7]).*

PhD Software Engineer Intern, **Google**, Sunnyvale, CA (Virtual). May 2020 – Aug. 2020
Managers: Bharath Krishna, Yili Ding
Developed configuration correctness review pipeline for load balancers.

Software Engineer Intern, **Credit Suisse**, Hong Kong. May 2018 – Aug. 2018
Manager: Sandeep Rakhra
Developed a data processing pipeline for risk information storage and analysis.

Software Engineer Intern, **Hututa Technologies**, Hong Kong. May 2017 – Aug. 2017
Managers: Dr. Zhiqiang Ma, Dr. Lin Gu
Developed a distributed data processing system to support biological gene mapping algorithms with networking monitoring for fault tolerance.

Conference & Journal Publications

Summary (**first-author**): **ATC [2024, 2023]**, **DSN** [**2024**, 2018], ICS 2021, ISCA 2024, **MLSys 2024**, NeurIPS [2023, 2022], NSDI 2018, **OSDI 2020**, **SoCC 2022**, **IEEE TSC 2022**

[C13] Power-aware Deep Learning Model Serving with μ-Serve
Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Saurabh Jha, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
In Proceedings of the 2024 USENIX Annual Technical Conference **(USENIX ATC 2024)**

[C12] FLASH: Fast Model Adaptation in ML-Centric Cloud Platforms
Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
In Proceedings of the 7th Annual Conference on Machine Learning and Systems **(MLSys 2024)**

[C11] When Green Computing Meets Performance and Resilience SLOs
Haoran Qiu, Weichao Mao, Saurabh Jha, Chen Wang, Hubertus Franke, Chandra Narayanaswami, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
In Proceedings of the 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks **(DSN 2024 Disrupt Track)**
Selected at NSF Workshop on Sustainable Computing for Sustainability 2024

[C10] AWARE: Automating Workload Autoscaling with Reinforcement Learning in Production Cloud
Haoran Qiu, Weichao Mao, Chen Wang, Alaa Youssef, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
In Proceedings of the 2023 USENIX Annual Technical Conference **(USENIX ATC 2023)**
Artifact Available, Functional, Reproduced
Selected for Presentation at KubeCon + CloudNativeCon NA 2023

[C9] SIMPPO: A Scalable and Adaptive Online Learning Framework for Multi-tenant Serverless Resource Management
Haoran Qiu, Weichao Mao, Archit Patke, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
In Proceedings of the 13th ACM Symposium on Cloud Computing **(SoCC 2022)**

[C8] FIRM: An Intelligent Fine-grained Resource Management Framework for SLO-oriented Microservices
Haoran Qiu, Subho S. Banerjee, Saurabh Jha, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer

In Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation **(USENIX OSDI 2020)**
Artifact Available, Functional, Reproduced
Dataset Selected as ICPE 2024 Data Challenge

[C7] SmartOClock: Workload- and Risk-Aware Overclocking in the Cloud
Jovan Stojkovic, Pulkit Misra, Íñigo Goiri, Sam Whitlock, Esha Choukse, Mayukh Das, Chetan Bansal, Jason Lee, Zoey Sun, **Haoran Qiu**, Reed Zimmermann, Savyasachi Samal, Brijesh Warrier, Ricardo Bianchini
In Proceedings of the 51th International Symposium on Computer Architecture **(ISCA 2024)**

[C6] Multi-Agent Meta-Reinforcement Learning: Sharper Convergence Rates with Task Similarity
Weichao Mao, **Haoran Qiu**, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer, Tamer Başar
In Proceedings of the 37th Conference on Neural Information Processing Systems **(NeurIPS 2023)**

[C5] A Mean-Field Game Approach to Cloud Resource Management with Function Approximation
Weichao Mao, **Haoran Qiu**, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer, Tamer Başar
In Proceedings of the 36th Conference on Neural Information Processing Systems **(NeurIPS 2022)**

[C4] Delay Sensitivity-driven Congestion Mitigation for HPC Systems
Archit Patke, Saurabh Jha, **Haoran Qiu**, Jim Brandt, Ann Gentile, Joe Greenseid, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
In Proceedings of the 35th ACM International Conference on Supercomputing **(ICS 2021)**

[C3] A Geography-Based P2P Overlay Network for Fast and Robust Blockchain Systems
Haoran Qiu, Tao Ji, Shixiong Zhao, Xusheng Chen, Ji Qi, Heming Cui, Sen Wang
In **IEEE Transactions on Services Computing**, vol. 16, no. 3, pp. 1572-1588, 2022
Best Undergraduate Thesis 2nd Runner-up

[C2] OWL: Understanding and Detecting Concurrency Attacks
Shixiong Zhao, Rui Gu, **Haoran Qiu**, Tsz On Li, Yuexuan Wang, Heming Cui, Junfeng Yang
In Proceedings of the 48th IEEE/IFIP International Conference on Dependable Systems and Networks **(DSN 2018)**

[C1] PLOVER: Fast, Multi-core Scalable Virtual Machine Fault-tolerance
Cheng Wang, Xusheng Chen, **Haoran Qiu**, Boxuan Li, Weiwei Jia, Shixiong Zhao, Heming Cui
In Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation **(USENIX NSDI 2018)**

Workshop Publications

[W9] Efficient Interactive LLM Serving with Proxy Model-based Sequence Length Prediction
Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
Workshop on Cloud Intelligence/AIOps @ASPLOS 2024

[W8] Efficient Queue Management for Large Language Model Serving
Archit Patke, Dhemath Reddy, Saurabh Jha, Christian Pinto, **Haoran Qiu**, Shengkun Cui, Chandra Narayanaswami, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
Workshop on Cloud Intelligence/AIOps @ASPLOS 2024

[W7] On the Promise and Challenges of Foundation Models for Cloud Systems Management
Haoran Qiu, Weichao Mao, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar,

Ravishankar K. Iyer
ML for Systems Workshop @NeurIPS 2023
Selected for Spotlight Presentation

[W6] PARM: Adaptive Resource Allocation for Datacenter Power Capping
Haoran Qiu, Linghao Zhang, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
ML for Systems Workshop @NeurIPS 2023

[W5] Reinforcement Learning for Resource Management in Multi-tenant Serverless Platforms
Haoran Qiu, Weichao Mao, Archit Patke, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer
EuroMLSys @EuroSys 2022

[W4] Is Function-as-a-Service a Good Fit for Latency-critical Services?
Haoran Qiu, Saurabh Jha, Subho S. Banerjee, Archit Patke, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
WoSC @Middleware 2021

[W3] Evaluating Hardware Memory Disaggregation Under Delay and Contention
Archit Patke, **Haoran Qiu**, Saurabh Jha, Srikumar Venugopal, Michele Gazzetti, Christian Pinto, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
COMPSYS @IPDPS 2022
Best Presentation Award

[W2] I-SPOT: Supporting Resilient Cloud Computing Using Counterfactual Reasoning
Harshitha Sreejith, Saurabh Jha, **Haoran Qiu**, Larisa Shwartz, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
IBM-IEEE CAS/EDS AICS 2022

[W1] Distributed System Design and Management with Inflight Analytics
Haoran Qiu, Saurabh Jha, Subho S. Banerjee, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
IBM-IEEE CAS/EDS AICS 2021

Selected Pre-prints

[P4] QLM: Queue Management for SLO-Oriented Large Language Model Serving (Under Review)
Archit Patke, Dhemath Reddy, **Haoran Qiu**, Saurabh Jha, Christian Pinto, Shengkun Cui, Chandra Narayanaswami, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer

[P3] INDIGO: Intelligent Page Migration for Hardware Memory Disaggregation (Under Review)
Archit Patke, Christian Pinto, **Haoran Qiu**, Shengkun Cui, Saurabh Jha, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer

[P2] Decision Transformer as a Foundation Model for Partially Observable Continuous Control (Under Review)
Xiangyuan Zhang, Weichao Mao, **Haoran Qiu**, Tamer Başar
arxiv.org/abs/2404.02407

[P1] SLO beyond the Hardware Isolation Limits
Haoran Qiu, Yongzhou Chen, Tianyin Xu, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer
arxiv.org/abs/2109.11666.

Invited Talks

ML-Centric Cloud Systems Management with Learned Abstractions
- Rice University (CS Department)

- University of Illinois Urbana-Champaign (CS Department PILOT Seminar)
- University of Connecticut (CSE Department)
- University of Hong Kong (CS Department)
- University of California, Santa Cruz (CSE Department)
- University of Illinois, Chicago (CS Department)

Power-aware Deep Learning Model Serving with μ-Serve
- CSL Student Conference 2024
- Hybrid Cloud and AI Workshop 2024, IBM-Illinois Discovery Accelerator Institute
- USENIX ATC 2024

Efficient and Robust Online Learning in Cloud Datacenters
- ML for Systems Seminar, Google Brain
- Huawei Cloud Research, Huawei Technologies
- Peisu Xia Young Scholars Forum 2022, Chinese Academy of Sciences
- MLSys Rising Star Workshop 2023
- UIUC Systems Research Seminar, 2024
- Microsoft Research Asia, 2024
- Google SRG (SystemsResearch@Google), 2024
- Sky Systems Lab @UC Berkeley, 2024

Towards a Foundation Model for Cloud Systems Management
- ML for Systems Workshop at NeurIPS 2023

Sustainable Workload Autoscaling with Predictive AI
- KubeCon/CloudNativeCon North America 2023, CNCF
- ML for Systems Workshop at NeurIPS 2023

Automate Workload Autoscaling in Production Cloud Systems with Meta-Learning
- USENIX ATC 2023
- AI Cloud Platform, IBM Research
- Hybrid Cloud and AI Workshop 2023, IBM-Illinois Discovery Accelerator Institute
- UIUC CS Systems and Networking Spring Retreat, 2023

Scalable and Adaptive Online Learning for Serverless Resource Management
- ACM SoCC 2022
- EuroMLSys 2022 (Co-located with ACM EuroSys 2022)
- Hybrid Cloud and AI Workshop 2022, IBM-Illinois Discovery Accelerator Institute

Intelligent Resource Management for SLO-oriented Cloud Microservices
- USENIX OSDI 2020
- AI Cloud Platform & Red Hat, IBM Research
- CSL Student Conference 2020
- Sustainable Hybrid Cloud Platform, IBM Research
- Turbonomics AIOps Research, IBM Research
- Applied AI Cloud, Infosys
- 2021 IBM IEEE CAS/EDS AI Compute Symposium

Teaching Experience

CS 536 Design of Reliable Systems and Networks, Course Assistant, UIUC	Spring 2024
CS 598 ML and Data Systems, Teaching Assistant, UIUC	Fall 2023
ECE 598 Dependable AI Systems, Course Assistant, UIUC	Fall 2023
ECE 471 Data Science Analytics, Course Assistant, UIUC	Fall 2022
CS 536 Design of Fault-Tolerant Digital Systems, Course Assistant, UIUC	Fall 2021

COMP 2396 Object-oriented Programming, Teaching Assistant, HKU Fall 2017

Professional Services

Program Committee / Reviewer

- Workshop on Cloud Intelligence and AIOps at ASPLOS 2024
- Journal of Systems Research (JSys), 2024
- ACM Transactions on Software Engineering and Methodology (TSEM), 2024
- IEEE Transactions on Automation Science and Engineering, 2024
- Journal of Network and Computer Applications, 2024
- Sustainable Computing: Informatics and Systems, 2024
- ML for Systems Workshop at NeurIPS 2023
- DSN 2023 Doctoral Forum
- ACM Transactions on Architecture and Code Optimization (TACO), 2023, 2024
- IEEE Internet of Things Journal, 2022
- International Journal of Numerical Modelling: Electronic Networks Devices and Fields, 2021
- Machine Learning Session, UIUC CSL Student Conference 2021

External Program Committee / Reviewer

- USENIX ATC 2024
- ACM EuroSys 2024
- ACM EuroSys 2022

Artifact Evaluation Committee

- 2023: SOSP, EuroSys, MLSys
- 2022: OSDI, ATC

Community Services

- Mentor, Promoting Undergraduate Research in Engineering (PURE), UIUC, 2023
- Mentor, Illinois Science and Technology Coalition (ISTC), 2023–24
- Member, Institute for Inclusion, Diversity, Equity and Access (IDEA), UIUC, 2023
- Mentor, Undergraduate Research Experience (URE) at IIDAI, 2022–23
- CS Ambassador, UIUC CS Visit Day for Prospective Graduate Students, UIUC, 2020
- Undergraduate Ambassador, HKU, 2016

References

References available upon request.