HANFEI YU

Email: hyu42@stevens.edu | Cell: (+1) 253-393-0977
Personal Web: hanfeiyu.github.io | Github: github.com/hanfeiyu

PERSONAL STATEMENT

I'm a fifth-year Ph.D. student in the Department of Electrical and Computer Engineering, at Stevens Institute of Technology, advised by Prof. Hao Wang. I received my Master's Degree in Computer Science and Systems at University of Washington Tacoma and my Bachelor's Degree in Electronic Engineering at Shanghai Jiao Tong University. I was a Research Intern at Microsoft M365 Research (Systems Innovation Research Group) in summer 2025, and a Research Intern at Microsoft Azure Research (Azure Research Systems) in summer 2024.

My research interests lie in **Serverless Computing**, **Large-Scale AI/ML Systems**, **LLM Serving Systems**, and **Reinforcement Learning Systems**. I develop efficient serverless AI eco-systems that integrate cloud and HPC resources to optimize large-scale AI workloads. I received the **ACM SoCC'24 Best Paper Award** and the recognition of the **ACM/IEEE SC'24 Best Student Paper Finalist**. I was selected as one of the **2025 MLCommons ML and Systems Rising Stars**.

EDUCATION

Stevens Institute of Technology, Hoboken, NJ, USA Sep 2024 - Present
Doctor of Philosophy in Computer Engineering

Louisiana State University, Baton Rouge, LA, USA June 2021 - Aug 2024
Doctor of Philosophy in Computer Science (transferred)

University of Washington, Tacoma, WA, USA Sep 2019 - Feb 2021
Master in Computer Science and Systems

Shanghai Jiao Tong University, Shanghai, China Sep 2015 - July 2019
Bachelor in Electronic Engineering

WORK EXPERIENCE

Stevens Institute of Technology, Hoboken, NJ, USA Sep 2024 - Present
Research Assistant and Teaching Assistant

· CPE 595 Applied Machine Learning

Microsoft M365 Research, Redmond, WA, USA May 2025 - Aug 2025
Research Intern at Systems Innovation Research Group

· Characterized production workloads of OpenAI Copilot services.
· Designed and implemented efficient KV cache management solutions for multimodal LLM serving.

Microsoft Azure Research, Redmond, WA, USA May 2024 - Aug 2024
Research Intern at Azure Research Systems

· Characterized production workloads of Azure container platforms.
· Designed new solutions to optimized resource efficiency of container orchestration.

Louisiana State University, Baton Rouge, USA June 2021 - Aug 2024
Research Assistant and Teaching Assistant

· CSC 4501 Computer Networks, CSC 3501 Computer Organization & Design, CSC 3102 Advanced Data Structures and Algorithms Analysis, CSC 2259 Discrete Structures, CSC 1350 Introduction to Computer Science

University of Washington, Tacoma, USA Sep 2020 - Jan 2021
Research Assistant and Teaching Assistant

· TCSS 305 Programming Practicum, TCSS 422 Operating Systems

Intel, Shanghai, China Aug 2018 - Feb 2019
Software Developer Intern at UEFI-BIOS Development Group

· Developed a cross-platform UEFI driver that analyzes network traffic with PCAP

PUBLICATIONS

[**ACM EuroSys'26**] Hanfei Yu, Xingqi Cui, Hong Zhang, Hao Wang, and Hao Wang. "Taming Latency-Memory Trade-Off in MoE-Based LLM Serving via Fine-Grained Expert Offloading." *The European Conference on Computer Systems.*

[**IEEE TPDS**] Yifan Sui, Hanfei Yu, Yitao Hu, Jianxun Li, and Hao Wang. "Accelerating ML Inference via Opportunistic Pre-Loading on Serverless Clusters." *IEEE Transactions on Parallel and Distributed Systems, 2025.*

[**ACM SoCC'25**] Rui Wei, Hanfei Yu, Xikang Song, Jian Li, Devesh Tiwari, Ying Mao, and Hao Wang. "Multi-Agent Reinforcement Learning with Serverless Computing." *ACM Symposium on Cloud Computing, 2025.*

[**arXiv'25**] Yifan Sui, Hao Wang, Hanfei Yu, Yitao Hu, Jianxun Li. "ServerlessLoRA: Minimizing Latency and Cost in Serverless Inference for LoRA-Based LLMs." *In-submission.*

[**VLDB'25**] Hanfei Yu, Jacob Carter, Hao Wang, Devesh Tiwari, Jian Li, Seung-Jong Park. "Nitro: Boosting Distributed Reinforcement Learning with Serverless Computing." *The International Conference on Very Large Data Bases, 2025.*

[**ACM SoCC'24, Best Paper Award**] Yifan Sui, Hanfei Yu, Yitao Hu, Jianxun Li, Hao Wang. "Pre-Warming is Not Enough: Accelerating Serverless Inference With Opportunistic Pre-Loading." *ACM Symposium on Cloud Computing, 2024.*

[**IEEE TPDS**] Hanfei Yu, Hao Wang, Jian Li, Xu Yuan, Seung-Jong Park. "Freyr+: Harvesting Idle Resources in Serverless Computing via Deep Reinforcement Learning." *IEEE Transactions on Parallel and Distributed Systems, 2024.*

[**ACM/IEEE SC'24, Best Student Paper Finalist**] Hanfei Yu, Hao Wang, Devesh Tiwari, Jian Li, Seung-Jong Park. "Stellaris: Staleness-Aware Distributed Reinforcement Learning with Serverless Computing." *ACM/IEEE The International Conference for High Performance Computing, Networking, Storage, and Analysis, 2024.*

[**AAAI'24**] Hanfei Yu, Jian Li, Yang Hua, Xu Yuan, Hao Wang. "Cheaper and Faster: Distributed Deep Reinforcement Learning with Serverless Computing." *Thirty-Eighth AAAI Conference on Artificial Intelligence, 2024.*

[**ACM ASPLOS'24**] Hanfei Yu, Rohan Basu Roy, Christian Fontenot, Devesh Tiwari, Jian Li, Hong Zhang, Hao Wang, Seung-Jong Park. "RainbowCake: Mitigating Cold-starts in Serverless with Layer-wise Container Caching and Sharing." *ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024.*

[**ACM HPDC'23**] Hanfei Yu, Christian Fontenot, Hao Wang, Jian Li, Xu Yuan, and Seung-Jong Park. "Libra: Harvesting Idle Resources Safely and Timely in Serverless Clusters." *ACM International Symposium on High-Performance Parallel and Distributed Computing, 2023.*

[**ACM WWW'22**] Hanfei Yu, Hao Wang, Jian Li, Xu Yuan, Seung-Jong Park. "Accelerating Serverless Computing by Harvesting Idle Resources." *ACM Web Conference, 2022.*

[**IEEE ACSOS'21**] Hanfei Yu, Athirai A. Irissappane, Hao Wang, Wes J. Lloyd. "FaaSRank: Learning to Schedule Functions in Serverless Platforms." *IEEE International Conference on Autonomic Computing and Self-Organizing Systems, 2021.*

[**ACM/SPEC ICPE'21**] Robert Cordingly, Navid Heydari, Hanfei Yu, Varik Hoang, Zohreh Sadeghi, Wes Lloyd. "Enhancing Observability of Serverless Computing with the Serverless Application Analytics Framework." *ACM/SPEC International Conference on Performance Engineering, 2021.*

[**WoSC 2020**] Robert Cordingly, Hanfei Yu, Varik Hoang, Zohreh Sadeghi, David Foster, David Perez, Rashad Hatchett, Wes Lloyd. "The Serverless Application Analytics Framework: Enabling Design Trade-off Evaluation for Serverless Software." *International Workshop on Serverless Computing, 2020.*

[**arXiv'20**] Athirai A. Irissappane, Hanfei Yu, Yankun Shen, Anubha Agrawal, Gray Stanton. "Leveraging GPT-2 for Classifying Spam Reviews with Limited Labeled Data via Adversarial Training."

[**IEEE CBDCom'20**] Robert Cordingly, Hanfei Yu, Varik Hoang, David Perez, David Foster, Zohreh Sadeghi, Rashad Hatchett, Wes J Lloyd. "Implications of Programming Language Selection for Serverless Data Processing Pipelines." *IEEE International Conference on Cloud and Big Data Computing, 2020.*

ACADEMIC SERVICES

— Conferences —

Ninth Annual Conference on Machine Learning and Systems (MLSys'26), Program Committee
International Conference on Learning Representations (ICLR'26), Reviewer
Fortieth AAAI Conference on Artificial Intelligence (AAAI'26), Program Committee
The European Conference on Computer Systems (EuroSys'26), Shadow Program Committee
USENIX Conference on File and Storage Technologies (FAST'26), Artifact Evaluation Program Committee
IEEE International Conference on Parallel and Distributed Systems (ICPADS'25), Technical Program Committee
ACM Symposium on Operating Systems Principles (SOSP'25), Artifact Evaluation Program Committee
ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT'25), Artifact Evaluation Program Committee
ACM International Conference on Mobile Systems, Applications, and Services (MobiSys'25), Artifact

Evaluation Program Committee
IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC'25), Program Committee
USENIX Conference on File and Storage Technologies (FAST'25), Artifact Evaluation Program Committee
International Conference on Learning Representations (ICLR'25), Reviewer
IEEE International Conference on Parallel and Distributed Systems (ICPADS'24), Technical Program Committee
ACM The Web Conference (WWW'24), Artifact Evaluation Program Committee
European Conference on Artificial Intelligence (ECAI'23), Reviewer
IEEE Global Communications Conference (GLOBECOM'22), Program Committee
EAI International Conference on Ad Hoc Networks (AdHocNets'21), Reviewer

— Journals —

Cluster Computing, Reviewer
IEEE Transactions on Services Computing (TSC), Reviewer
The Journal of Supercomputing (TJSC), Reviewer
ACM Transactions on Autonomous and Adaptive Systems (TAAS), Reviewer
Performance Evaluation (PEVA), Reviewer
IEEE Transactions on Computers (TC), Reviewer
IEEE Transactions on Mobile Computing (TMC), Reviewer
IEEE Transactions on Parallel and Distributed Systems (TPDS), Reviewer
IEEE Internet of Things Journal (IoTJ), Reviewer
IEEE Transactions on Network Science and Engineering (TNSE), Reviewer
Journal of Systems Architecture (JSA), Reviewer
IEEE Transactions on Cloud Computing (TCC), Reviewer

STUDENT MENTORING

Xingqi Cui, PhD student at Rice University, second-author paper in EuroSys'26
Jacob Carter, PhD student at University of Florida, second-author paper in VLDB'25
Christian Fontenot, PhD student at University of Colorado Boulder, third-author paper in ACM ASPLOS'24 and second-author paper in ACM HPDC'23