HANFEI YU

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

RESEARCH INTERESTS

My research interests lie in **Large-Scale AI/ML Systems**, **LLM serving Systems**, **Reinforcement Learning Systems**, **Serverless Computing**, and **Cloud Computing**. Specifically, I focus on improving the resource efficiency of serverless computing systems with AI/ML-driven techniques and building efficient serverless systems for AI/ML training and inference. I'm the recipient of various academic awards, including SoCC'24 Best Paper Award, SC'24 Best Student Paper Finalist, and ECAI'23 Call to Arms Award.

EDUCATION

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

Louisiana State University, Baton Rouge, LA, USA Sep 2021 - Aug 2024
Doctor of Philosophy in Computer Science (transferred out)
GPA: 3.72/4.0

University of Washington, Tacoma, WA, USA Sep 2019 - Feb 2021
Master in Computer Science and Systems
GPA: 3.96/4.0

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

WORK EXPERIENCE

Stevens Institute of Technology, Hoboken, NJ, USA Sep 2021 - Present
Research Assistant

· IntelliSys Lab

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

· Characterized production workloads of Azure Container Instances and Azure Kubernetes Services.
· Optimized resource efficiency of container orchestration by designing new solutions.

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

· IntelliSys Lab

Louisiana State University, Baton Rouge, USA Jan 2020 - June 2020
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

· Cloud and Distributed Systems (CDS) Research Lab

University of Washington, Tacoma, USA Jan 2020 - June 2020
Teaching Assistant

· TCSS 305 Programming Practicum, TCSS 422 Operating Systems

Intel, Shanghai, China Aug 2018 - Feb 2019
Asia-Pacific Research & Development Ltd
Software Developer Intern, UEFI-BIOS Development Group

· Contributed to the development of a UEFI Driver, which works cross-platform on Windows and Linux. The driver filters, extracts, and analyzes network packets by leveraging an open-source library called PCAP. I also helped implement and test new features of the driver.
· Wrote Shell and Python scripts to enable automatic driver installation, dependency collection, and product testing using internal tools.

PUBLICATIONS

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

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

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

Stellaris: Staleness-Aware Distributed Reinforcement Learning with Serverless Computing
Hanfei Yu, Hao Wang, Devesh Tiwari, Jian Li, Seung-Jong Park
The International Conference on Very Large Data Bases (**SC 2024, Best Student Paper Finalist**)

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

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

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

Accelerating Serverless Computing by Harvesting Idle Resources
Hanfei Yu, Hao Wang, Jian Li, Xu Yuan, Seung-Jong Park

ACM Web Conference (**WWW 2022**)

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

FaaSRank: A Reinforcement Learning Scheduler for Serverless Function-as-a-Service Platforms
Hanfei Yu
Master Thesis, University of Washington

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

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

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

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

ACADEMIC SERVICES

2024
USENIX Conference on File and Storage Technologies (FAST), Artifact Evaluation Program Committee
International Conference on Learning Representations (ICLR), Reviewer
IEEE International Conference on Parallel and Distributed Systems (ICPADS), Technical Program Committee
ACM The Web Conference (WWW), Artifact Evaluation Program Committee
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
2023
Journal of Systems Architecture (JSA), Reviewer IEEE Transactions on Cloud Computing (TCC), Reviewer
IEEE Global Communications Conference (GLOBECOM), Reviewer
European Conference on Artificial Intelligence (ECAI), Reviewer
IEEE Transactions on Parallel and Distributed Systems (TPDS), Reviewer
2021
EAI International Conference on Ad Hoc Networks (AdHocNets), Reviewer