Yuanli Wang

(+1) 612-963-5189 | yuanliw@bu.edu | https://pentium3.github.io/

SUMMARY

Fifth year PhD student at Boston University. My research interests are in **distributed systems** and **stream processing systems**. I also work on **Agentic AI** systems and applications.

EDUCATION

Boston University Boston, MA
PhD in Computer Science *09/2021 – 12/2026 (Expected)*
- Advisor: Vasiliki Kalavri

University of Minnesota, Twin Cities Minneapolis, MN
Master of Science in Computer Science *09/2018 – 06/2021*

Hefei University of Technology Anhui, China
Bachelor of Science in Computer Science *09/2013 – 06/2017*

RESEARCH EXPERIENCE

Scaling Agentic AI Workflows on Unified Memory Workstations *01/2026 – Present*
- Designing a bandwidth-aware scheduler and auto-configuration framework for multi-agent systems running on unified-memory workstations.
- Developed and evaluated SSD-based KV-cache offloading for vLLM on NVIDIA DGX Spark using GPU-Direct Storage.

WebAgent Behavior Analysis and Visualization via Graph-Based Evaluation *05/2025 – Present*
- Designed a graph-based evaluation framework that aggregates trajectories from multiple webagent frameworks into per-task weighted action graphs via LLM-based annotation, canonicalization, and deterministic graph merging algorithms, to analyze trajectories beyond binary success metrics.
- Validated the proposed framework across 4768 trajectories spanning 812 webagent tasks. Designed multiple reward mechanisms to expose redundancy/inefficiency and discover critical actions overlooked by outcome-only metrics.

Disaggregated and Self-Managed Data Stream Processing System *09/2021 – Present*
- Working on a novel fully disaggregated architecture for data stream processing systems that completely decouples state management from computation, enabling zero-downtime reconfiguration.
- Proposed *CAPSys*, an adaptive resource controller for dataflow stream processors, that considers auto-scaling and task placement in concert. Achieved orders of magnitude lower computing time and up to 6× higher throughput with fixed resources, compared to the state-of-the-art work.
- Proposed an adaptive edge–cloud stream processing system compatible with Apache Flink that rewrites queries into fault-tolerant segments and dynamically reroutes computation without downtime.

SELECTED PUBLICATIONS

[1]. CAPSys: Contention-aware task placement for data stream processing. Yuanli Wang, Lei Huang, Zikun Wang, Vasiliki Kalavri, and Abraham Matta. Proceedings of the Twentieth European Conference on Computer Systems (EuroSys 2025). 2025

[2]. WebGraphEval: Multi-Turn Trajectory Evaluation for Web Agents using Graph Representation. Yaoyao Qian, Yuanli Wang, Jinda Zhang, Yun Zong, Meixu Chen, Hanhan Zhou, Jindan Huang, Yifan Zeng, Xinyu Hu, Chan Hee Song, Danqing Zhang. NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models.

[3]. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. ICLR 2026.

[4]. The *Non-Expert Tax*: Quantifying the cost of auto-scaling in Cloud-based data stream analytics. Yuanli Wang, Baiqing Lyu, Vasiliki Kalavri. International Workshop on Big Data in Emergent Distributed Environments (BiDEDE 2022) , co-located with SIGMOD'22. 2022

[5]. A New Benchmark Harness for Systematic and Robust Evaluation of Streaming State Stores. Esmail Asyabi, Yuanli Wang, John Liagouris, Vasiliki Kalavri, Azer Bestavros. Proceedings of the Seventeenth European Conference on Computer Systems (EuroSys 2022). 2022

[6]. HACCS: Heterogeneity-Aware Clustered Client Selection for Accelerated Federated Learning. Joel Wolfrath, Nikhil Sreekumar, Dhruv Kumar, Yuanli Wang, Abhishek Chandra. 36th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2022). 2022

[7]. Accelerated Training via Device Similarity in Federated Learning. Yuanli Wang, Joel Wolfrath, Nikhil Sreekumar, Dhruv Kumar, Abhishek Chandra. The 4th International Workshop on Edge Systems, Analytics and Networking (EdgeSys 2021). 2021

[8]. Fail-slow fault tolerance needs programming support. Andrew Yoo, Yuanli Wang, Ritesh Sinha, Shuai Mu, Tianyin Xu. The 18th Workshop on Hot Topics in Operating Systems (HotOS XVIII). 2021

PROFESSIONAL EXPERIENCE

Megagon Labs | *Research Intern* 05/2025 – 08/2025

- Designing a multi-agent system for decision-making under uncertainty , which uses LLMs and probabilistic inference to automatically discover and model causal relationships between hidden factors from data.
- Preparing a first-author research paper based on this study.

Apple | *Data Processing Platform Intern* 05/2023 – 08/2023

- Integrated OpenLineage framework with Flink to track data lineage for Apple's AIML data processing platform. Implemented new OpenLineage visitors for Kafka and Iceberg data connectors in Flink.
- Contributed to OpenLineage open source repository.

PingCAP | *Database Engineer Intern* 05/2019 – 08/2019

- Worked on AutoTiKV project from scratch: used machine learning to tune a database system under user-specific workloads.
- Implemented a Gaussian Process Regression Model to predict the performance of RocksDB(the core storage engine in TiKV) under different knob configurations. Achieved 1.3x lower latency under several types of workloads without human guidance.

PROFESSIONAL SERVICE

- **Reviewer:** ICLR 2026, ICLR 2025, ICASSP 2025, IJCNN 2025 , NAACL 2025 (Demo Track) , IJCAI 2025 (Demo Track), Future Generation Computer Systems, Internet of Things Journal

SKILLS

Programming: Python, Java, C++, Go, Rust, Shell, SQL, Docker, Git, Flink, Beam, Kafka, RocksDB

Cloud infrastructures: AWS, Google Cloud, Microsoft Azure

SELECTED AWARDS

Rank 16/183 in 2018 ACM-ICPC North Central North America Regional Contest 11/2018

Bronze Medal, China Collegiate Programming Contest 10/2015