Yuan Yao

University of Southern California | +1 (734) 882-9251 | yyao3787@usc.edu | yuanyao.me

Education

University of Southern California, Los Angeles **California, USA**
Ph.D. in Computer Science (Aug. 2023 -) *GPA: 4.00 / 4.00*

University of Michigan, Ann Arbor **Michigan, USA**
Ph.D. in Computer Science and Engineering (Aug. 2021 - Aug. 2023) *GPA: 4.00 / 4.00*
Advised by: *Professor Harsha V. Madhyastha*

University of Michigan, Ann Arbor **Michigan, USA**
B.Eng. in Computer Science and Engineering (Aug. 2019 - May. 2021) *GPA: 3.95 / 4.00*

Shanghai Jiao Tong University UM-SJTU Joint Institute **Shanghai, China**
B.S.E in Electrical and Computer Engineering (Sept. 2017 - Aug. 2021) *GPA: 3.70 / 4.00*

Skills: C++, Rust, Java, Go, Python / Google Cloud, AWS / Linux, Docker, Kubernetes, Flask, TensorFlow / Git, LaTeX
Areas of Interest: Distributed systems / Cloud computing / Machine Learning systems

Project Experience

Cost-Effective Support for Cloud-Assisted 3D Printing **Los Angeles, CA**
Graduate Student Research Assistant, advised by Prof. Harsha Madhyastha *Aug. 2021 - Jan. 2025*

- Independently developed *Cosmic*, a serverless framework optimizing cloud-assisted control of 3D printing.
- Leveraged AWS Lambda to reduce idle cost and over-provisioning, cutting computation costs by 2.8× compared to VM-based solutions.
- Implemented speculative execution and group partitioning to achieve stringent millisecond-level timing requirements, ensuring timely execution across diverse print workloads.
- Collaborated with a 3D printing research team as project lead, validating framework feasibility; independently authored and submitted research paper, accepted by USENIX ATC'25.

Towards Sub-second Serverless Serving of Large Models **Los Angeles, CA**
Graduate Student Research Assistant, advised by Prof. Harsha Madhyastha *Sept. 2024 - May 2025*

- Benchmarked and analyzed model serving latency across various serverless platforms, identifying critical bottlenecks through detailed latency visualization.
- Developed a prototype tool optimizing inference initialization by reducing framework dependencies, decreasing startup latency; designing experiments toward communication-latency-aware optimization for larger-scale distributed GPU inference.

Cloud-Accelerated Real-Time Telepresence System **Los Angeles, CA**
Doctoral-level Computer Networks Course Project, advised by Prof. Ramesh Govindan *Jan. 2024 - May. 2024*

- Led development of a real-time 3D telepresence system using cloud GPUs and deep learning for webcam-based human rendering.
- Deployed on Azure with NVIDIA V100 GPUs, integrating EasyMocap, PPHumanSeg, and GauHuman models.
- Ensured smooth AR/VR experience by analyzing and minimizing network and processing delays.
- Achieved <150 ms latency and 30 FPS by optimizing inference with batching and image quality tuning.

Consistency Analysis of Data Usage Purposes in Mobile Apps **Ann Arbor, MI**
Research Assistant at Real-Time Computing Laboratory, directed by Prof. Kang Shin *May 2020 - Sept. 2021*

- Single-handedly developed a smartphone debugging framework capturing and analyzing over 2 million data traffic instances from 20,000+ Android apps, identifying inconsistencies in 15% of apps, contributing to privacy policy enhancements.
- Independently built a web server supporting crawling for 1M+ app IDs and a webserver to monitor and analyze the captured data.
- Collaborated with privacy experts to annotate data, train models, and co-authored a published research paper.

Connecting high-resolution 3D chromatin organization with epigenomics **Ann Arbor, MI**
Research Assistant at Liu Lab, directed by Prof. Jie Liu *Jan. 2020 - Mar. 2022*

- Collaborated with team to develop a data pipeline to collect, preprocess, and impute epigenomic data, incorporating a deep learning model to map epigenomic features to 3D chromatin organization.

- Solely responsible for integrating the pipeline into the Chromosomal Structure And Epigenomics Analyzer (CAESAR), a web system built using Python Flask; the system hosts and visualizes data via the Nucleome Browser framework, allowing real-time user interaction for attribution calculation; collaborated on manuscript writing and publication.

Publications

- Cosmic: Cost-Effective Support for Cloud-Assisted 3D Printing
  Yuan Yao*, Chuan He, Chinedum Okwudire, Harsha V. Madhyastha*
  (To appear in) 2025 USENIX Annual Technical Conference
- Connecting high-resolution 3D chromatin organization with epigenomics [page]
  Fan Feng, **Yuan Yao***, Xue Qing David Wang, Xiaotian Zhang, Jie Liu*
  2022 Nature Communications
- Consistency Analysis of Data Usage Purposes in Mobile Apps [page] [pdf]
  Duc Bui, **Yuan Yao***, Jongmin Choi, Junbum Shin, Kang G. Shin*
  2021 ACM SIGSAC Conference on Computer and Communications Security (CCS ’21)