Yuan Yao

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

Summary

Master's student in Computer Science with 3+ years of research engineering experience building scalable backend systems and cloud-native applications. Strong foundation in distributed systems, cloud infrastructure, and machine learning workflows. Published at USENIX ATC and ACM CCS. Led research-driven engineering projects from prototype to production.

Education

University of Southern California, Los Angeles — **California, USA**
MS in Computer Science (Aug 2023 – Expected May 2026) — *GPA: 4.00 / 4.00*

University of Michigan, Ann Arbor — **Michigan, USA**
MSE in Computer Science and Engineering (Aug 2021 – Aug 2023) — *GPA: 4.00 / 4.00*
BSE in Computer Science and Engineering (Aug 2019 – May 2021, dual degree with SJTU) — *GPA: 3.95 / 4.00*

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

Work Experience

Research Assistant – Distributed & Cloud Systems — **Ann Arbor, MI / Los Angeles, CA**
Paid research engineer role; University of Michigan, University of Southern California — *Aug 2021 – May 2025*

- Built distributed systems, serverless platforms, and ML pipelines on AWS, Azure, and Google Cloud.
- Delivered low-latency, cost-efficient systems by improving cold starts, autoscaling, and data paths.
- Launched and operated services with hardware teams and cloud providers; adopted by internal research teams.

Teaching Assistant – Advanced Operating Systems, Introduction to Operating Systems — **Los Angeles, CA**
CSCI 555/655, CSCI 350; University of Southern California — *Spring 2024, Summer 2025*

- Assisted instruction, mentored debugging/system design, created and graded programming assignments and exams.

Project Experience

Cost-Effective Support for Cloud-Assisted 3D Printing — **Los Angeles, CA**
Project Lead, Graduate Research Assistant, advised by Prof. Harsha Madhyastha — *Aug 2021 – July 2025*

- Developed *Cosmic*, a serverless framework optimizing cloud-assisted control of 3D printing.
- Leveraged AWS Lambda to cut computation costs by 2.8× compared to VM-based solutions.
- Designed and implemented speculative execution and group partitioning to achieve stringent millisecond-level timing requirements, ensuring timely execution across diverse print workloads.
- Led interdisciplinary collaboration with a 3D printing research team; independently authored and presented paper at USENIX ATC '25.

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

- Initiated research on cold-start latency optimization in serverless GPU inference.
- Built benchmarking and latency profiling pipelines across platforms to identify bottlenecks.
- Prototyped solutions reducing framework overhead; currently designing communication-aware optimizations for large-scale distributed inference.

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

- Built a real-time 3D telepresence system using cloud-based deep learning for human rendering from webcam inputs.
- Deployed on Azure with NVIDIA V100 GPUs; integrated pose estimation and mesh generation models.
- Ensured smooth AR/VR experience by analyzing and minimizing network and processing delays.
- Achieved <150 ms end-to-end latency and 30 FPS through inference batching and adaptive 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*

- Designed a debugging framework that captured and analyzed 2M+ data traffic instances from 20K+ Android apps across multiple smartphones.
- Identified inconsistencies in 15% of apps, informing privacy policy improvements.
- Built a web crawler for over 1M app IDs and a monitoring server to track and analyze captured data streams.
- Collaborated with privacy experts to annotate data, train models, and co-authored peer-reviewed ACM CCS '21 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*

- Developed a data pipeline to collect, preprocess, and impute epigenomic data, incorporating a deep learning model to map epigenomic features to 3D chromatin organization.
- Integrated the pipeline into CAESAR, a web-based chromatin analysis system built with Python Flask; enabled real-time, interactive visualization.
- Sole system integrator; contributed to Nature Communications publication.

Technical Skills

Languages:	C++, Java, Python, Rust, SQL	**Cloud Platforms:**	AWS, Azure, Google Cloud
Infrastructure:	Linux, Docker, Kubernetes	**Frameworks/ML:**	Flask, PyTorch, TensorFlow
Tools:	Git, LaTeX		

Publications

- Cosmic: Cost-Effective Support for Cloud-Assisted 3D Printing [page]
  ***Yuan Yao**, Chuan He, Chinedum Okwudire, Harsha V. Madhyastha*
  2025 USENIX Annual Technical Conference (ATC '25)
- GenomicKB: a knowledge graph for the human genome [page]
  *Fan Feng, Feitong Tang, Yijia Gao, Dongyu Zhu, Tianjun Li, Shuyuan Yang, **Yuan Yao**, Yuanhao Huang, Jie Liu*
  2023 Nucleic Acids Research
- 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]
  *Duc Bui, **Yuan Yao**, Jongmin Choi, Junbum Shin, Kang G. Shin*
  2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)