Haocheng (Harvey) Yuan

haochengyhc@gmail.com | +1 (858) 396-4890 | linkedin.com/in/haocheng-yuan-harvey/

Education

University of California, San Diego (UCSD) *Master of Science in Computer Science* 09/2024 – 03/2026 (Expected)
University of Nottingham *Bachelor of Science in Computer Science* **GPA:** 3.94/4.0, Top 5% 09/2020 – 06/2024

Skills

Programming Languages: C++, Python, Java
Systems Performance: Multithreading, Async Processing, Memory Management, Latency Optimization, Distributed Systems
AI / ML Systems: Model Serving, Inference Pipelines, LLM Workloads, Performance Modeling & Benchmarking
Frameworks: PyTorch, TensorFlow, gRPC, FastAPI, Spring Boot
Infrastructure: AWS, Docker, Kubernetes, Redis, CI/CD

Work Experience

Software Development Engineer Intern 6/2025 - 9/2025
Amazon Web Services, Key Management Service Team

- Implemented low-latency request handling mechanisms in a distributed system operating at millions of requests per second, reducing P99 latency by **14%** and P100 latency by **64%** at 5M TPS.
- Conducted performance analysis and benchmarking to identify bottlenecks and guide system-level optimizations.
- Implemented traffic control and rate-limit throttling algorithm to improve system stability under bursty workloads.
- Collaborated with senior engineers on production rollout, monitoring, and rollback for a mission-critical cloud service.

Software Development Engineer Intern 6/2021 - 12/2021
IceWould, Software Development Team

- Developed backend services to support real-time AR workloads, focusing on request handling and system responsiveness.
- Built a responsive React frontend and integrated third-party SDKs via well-defined service interfaces.
- Implemented automated testing and CI/CD pipelines to improve release quality and reduce regression issues.

Selected Project

Distributed LLM Inference Benchmark & Optimization 10/2025 - present

- Developed a **C++** distributed LLM inference system using LibTorch and gRPC, supporting multi-node parallel inference.
- Implemented async and multi-threaded request handling, improving throughput by 2.5× and reducing P99 latency by 18%.
- Conducted end-to-end benchmarking and performance modeling, analyzing GPU utilization and latency bottlenecks under varying batch sizes and concurrency.
- Designed load balancing strategies across nodes, demonstrating scalable distributed AI workload optimization.

Travel Visa Assistant App with LLM 09/2025 – 10/2025

- Developed an **AI-driven** web application for visa preparation, integrating the OpenAI API to assist users with document requirements, eligibility, and application steps.
- Implemented a Redis-powered RAG pipeline to deliver accurate visa information and avoid LLM hallucinations.
- Built a conversational destination-recommendation chat interface to enhance usability.

Scheduling and Optimization in Large Ports 08/2023 – 06/2024

- Developed a reinforcement learning system, improving vehicle dispatch efficiency by **34%** at a major global port.
- Designed and built a distributed training pipeline leveraging socket and RPC, to enable communication between machines for **parallelize computation** and improve solution inference speed by **72%**.
- Designed a hyper-heuristic evolutionary algorithm and surrogate model to address sparse reward issues.

Publications

- **H. Yuan**, X. Chen, J. Zhu and R. Bai, “A Simulation Hyper-Heuristic Method for Multi-Floor AGV Delivery Services in Hospitals,” 2023 IEEE Symposium Series on Computational Intelligence