Hayes Zhao

hayes.zhao09@gmail.com • 703-915-8580 • www.linkedin.com/in/hayes-zhao • https://www.kaggle.com/hayeszhao

Computer Engineer focused on AI infrastructure, distributed computing, and model evaluation. Experienced building large-scale data pipelines, benchmarking and validation pipelines for deep learning models (including LLMs), and optimization systems that improve infra utilization and reduce operational overhead.

Experience

Data Engineer **| Bytedance** Ashburn, VA | Sep. 2025 – Present

Global ETL Pipeline (Business and Finance Data base)

- Built a production ETL pipeline integrating business and finance databases for global Data Center operations, improving consistency from 98.0% to 99.9% and ensuring reliability through robust data processing.
- Orchestrated daily ETL jobs with Apache Airflow using Python scripts to process 1 M assets and 10 K incremental updates, implementing retry logic and SLA monitoring that maintained near-real-time data availability.
- Scaled transformations and reconciliation logic with Spark, reducing manual triage by 90% through automated validation checks and improving accuracy in data center operations.
- Optimized data processing workflows to enhance Infrastructure performance and reliability.

Capacity & Compute Planning

- Developed a Python-based planning system combining a MILP solver with reinforcement learning policies to optimize procurement, placement, and migration under power and lead-time constraints, delivering faster plan generation and reducing overall procurement cost.
- Modeled downtime and migration costs for an internal fleet, selecting plans maximizing ROI while meeting availability requirements.
- Delivered +6% average compute uplift at fixed power via optimized allocation and migration schedules; built offline evaluation and guardrails for reliability.

Research Assistant **| George Washington University** Washington, DC | Mar. 2024 – Sep. 2024

- Built a topology-aware reinforcement learning framework with a hybrid GNN-Transformer mapper, boosting multi-accelerator inference throughput by 26.1% through advanced AI/ML techniques and research in artificial intelligence.
- Extended a C++ compiler toolchain to deploy 30+ DL models (LLM/CV/MoE) as computation graphs on heterogeneous accelerators, improving deployment efficiency and accuracy using programming skills.
- Reduced graph size by up to 90% via modularity-based clustering and co-authored STARDUST (IEEE CAL), cutting sample requirements by 15×, demonstrating strong research and accuracy in AI/ML.

Software Engineer Intern **| DHC Software** Beijing, CN | Jul. 2022 – Dec. 2022

- Designed a high-concurrency sales system to handle traffic spikes, boosting throughput by 100× (peak QPS from 500 to 50k+) by introducing Redis for hot-data caching and RocketMQ for asynchronous traffic leveling.
- Transitioned a monolithic backend into Java microservices (Spring Boot / Spring Cloud) with service discovery, API gateway routing, and fault-tolerant middleware, improving system resilience and scalability.
- Enhanced database performance through sharding and read–write separation, reducing latency under large-scale concurrent workloads.

Skills & Interests

Languages: Python, C++, Java, SQL, CUDA

ML/Systems: Reinforcement Learning, Optimization (MILP), GNN/Transformers, Compiler/Graph IR, Algorithms, Data Structures, Information Retrieval, Artificial Intelligence, Natural Language Processing,PyTorch, DeepSpeed,vLLM, AI/ML

Infra: Apache Airflow, Spark, Docker, Kubernetes, Redis, RocketMQ, Data Center, Distributed Computing, google cloud

Recognition: Kaggle Master (with 2 silver medals in natural language processing and recommendation system)

Education

MS. Computer Engineering | George Washington University Washington,DC | 2023 - 2025