Hayes Zhao

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

*Computer Engineer (M.S.) with a proven track record of accelerating **distributed high-performance systems** and **AI infrastructure**, delivering quantifiable industry performance gains and publishing research in the **top-tier IEEE Computer Architecture Letters (CAL)***

Education

MS. Computer Engineering | George Washington University Washington,DC | 2023 - 2025
*Focus on AI infrastructure. Awarded Graduate Research Assistant. Scholar. **Nominated to** **Sigma Xi**.*

BS. Electrical Engineering | Tianjin University of Technology and Education Tianjin, CN | 2019 - 2023
Focus and machine learning applications. Awarded in CICSIC.

Experience

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

- **Accelerated multi-accelerator inference throughput by 26.1%** by developing a topology-aware **Reinforcement Learning** framework featuring a hybrid **GNN-Transformer** architecture.
- Optimized a custom **C++ compiler** to partition and deploy 30+ diverse **large-scale deep learning models** (**LLM**, CV, MoE) across scalable, heterogeneous accelerator systems.
- **Co-first-authored** 'STARDUST' (IEEE Computer Architecture Letters), a Reinforcement Learning framework that **reduced training sample requirements** by **15x** compared to baseline methods.
- Architected the core two-stage optimization in STARDUST: first reducing computational graph size by up to **90%** via **Deep Modularity Network(DMN) clustering**, then applying a Reinforcement Learning (PPO) agent for highly efficient mapping.

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

- **Scaled backend system throughput by 100×** during high-traffic sales by implementing **distributed caching** (Redis) and **asynchronous messaging** (RocketMQ).
- Modernized a **monolithic architecture** by designing **microservices** using Java-based frameworks (Spring Boot, Spring Cloud) and Domain-Driven Design, reducing system complexity and operational costs.
- Built robust transaction systems with **distributed locking** and **rollback protocols** (TCC), ensuring consistency under load.
- Improved system reliability by developing rate-limiting and **abuse-prevention features**, protecting high-demand services during peak usage.

Projects

Distributed Key-Value storage system

- Built a distributed key-value store with **99.9% availability** and **1M QPS** leveraging consistent hashing for balanced sharding.
- **Reduced read latency by 42%** via async Apply, ReadIndex, and FollowerRead for non-blocking Raft-based reads.
- Designed **scalable data migration** and load balancing strategies across Raft groups.
- Optimized storage using RocksDB, B+ trees, hash tables, and MVCC for concurrency control.

[Kaggle Contest]H&M Personalized Fashion Recommendations (Ranked 75th/3759, Silver medal)

- **Boosted recommendation quality (MAP@12) by 3–5%** for a system personalizing a 106K-item catalog, achieved through advanced feature engineering and a robust model ensemble strategy.
- Engineered a **multi-stage ranking pipeline**, utilizing co-purchase heuristics and a **two-tower model** for candidate generation, followed by a **fine-tuned LightGBM** ensemble for final ranking.
- Developed rich temporal features (e.g., purchase frequency over 1/4/8-week windows) and **addressed data imbalance** with **negative sampling** to significantly improve model performance and robustness.

Awards & Honors

Associate Membership - Sigma Xi, The Scientific Research Honor Society. Inducted July 2025

Skills & Interests

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

AI tools: PyTorch, TensorFlow, DeepSpeed, NeMo, JAX, VLLM, TensorRT

Infrastructure & Tools: Redis, RocketMQ, RabbitMQ, TravisCI, Spring Boot/Cloud, DynamoDB, Node.js, Flask, FastAPI, Docker, Kubernetes, Git