

RUI HU

(412) 551-0300, ruihu.xyz@gmail.com
github.com/mrhoray

EXPERIENCE

Software Engineer (Staff, Sr. Staff) at DoorDash

October 2020 – Present

Org TL for Consumer domain (12 teams, 140+ ICs); focused on ML perf & infra since H2 2025

- 20x+ training speedup with 90%+ cost reduction via model, data-loading, and training-recipe optimizations
- 90%+ inference latency reduction with 95%+ cost savings via model and GPU serving optimizations
- Annualized \$80M+ GMV, \$20M+ variable profit, and \$600K+ infra savings from 4 system optimizations
- Annualized \$130M+ GMV from GenAI products, enabled by embedding-based retrieval at 1ms p90 / 4ms p999
- Increased experiments/eng/quarter by 89% via impact/effort ROI analysis and bottom-up rapid experiments
- Reduced time to start E2E testing by ~90% (20–30 min → 1–3 min/change), reaching 200+ monthly active developers and 3.3K+ test sessions/month in the launch quarter, via bidirectional local↔prod routing
- 1K+ stale feature flags and 140K+ LOC removed via data & code analysis, leaderboard, and area advocates

Software Engineer (Staff) at Uber Freight

October 2017 – September 2020

- Built the Freight Marketplace stack from the ground up, powering and optimizing a \$1B run-rate business
- Formed core technical domains and served as TL for Freight Marketplace (45+ ICs)
- Reduced MySQL primary load by 90%+ via contextual routing while preserving read-after-write consistency

Software Engineer (I, II, Senior, Staff) at Uber Marketplace Platform

May 2014 – September 2017

- As part of an 8-person tiger team, re-architected dispatch and fulfillment systems enabling 30x+ trip growth
- Built an ingress routing gateway enabling city-based sharding and failover of Marketplace systems, serving 10M+ trips/day at 160K+ RPS, 2ms p50 / 10ms p99
- Cut Marketplace compute 15%+ and mobile bandwidth 20%+ by replacing mobile short-polling with push

OPEN SOURCE PROJECTS

zero-to-hero

github.com/mrhoray/zero-to-hero

- From-scratch series on GPU programming, neural networks, reinforcement learning, and LLM inference
- Covers autograd, GPT/GPT-2, GPU kernels in Mojo, and tabular, deep, and planning-based RL
- WIP LLM inference engine inspired by nano-vLLM, mini-SGLang, and TokenSpeed, approaching SOTA perf

spar

github.com/mrhoray/spar

- Structured, parallel automated research combining Monte Carlo tree search with coding agents
- Outperforms a native-agent baseline with fewer candidates and lower variance

kdtree-rs

github.com/mrhoray/kdtree-rs

- K-dimensional tree for fast geospatial indexing and nearest-neighbor lookup
- High-performance implementation in pure, safe Rust with 20+ contributors

crc-rs

github.com/mrhoray/crc-rs

- Rust implementation of multi-width CRC supporting a broad range of standards
- Widely adopted CRC library in the Rust ecosystem, with 260M+ downloads

torch

github.com/mrhoray/torch

- CPU profiling tool for generating FlameGraphs from perf samples

swim-js

github.com/mrhoray/swim-js

- SWIM decentralized cluster membership in JavaScript for service discovery and distributed systems

EDUCATION

Carnegie Mellon University, Pittsburgh, PA

2012–2014

Southeast University, Nanjing, Jiangsu, China

2008–2012