Nathan Wang

(408) 829-2546 | nathan.r.wang@gmail.com | linkedin/nathan-r-wang | github/thecodingwizard | thecodingwizard.me

Education

Massachusetts Institute of Technology May 2026
B.S. in Computer Science *GPA: 5.0/5.0*
Coursework: Distributed Systems, Operating Systems, Software Performance, Machine Learning, Advanced Algorithms

Awards & Accomplishments

4x USA Computing Olympiad (USACO) Finalist 2019 – 2022

- One of 26 students invited to the United States Computing Olympiad’s national training camp.
- Spent 1000+ hours writing 100,000+ lines of C++ code to solve challenging algorithmic problems.

5x American Invitational Mathematics Exam (AIME) Qualifier 2017 – 2021

Experience

Member of Technical Staff | Modal Sep 2024 – Present

- Modal is a serverless cloud computing platform. I interned in January 2024 and am currently taking a gap year to work as a full-time software engineer.
- Built a real-time I/O streaming system in Rust for command-line access to running containers and interactive debugging of Python functions, improving developer experience.
- Investigated networking issues, reducing the latency of our serverless web endpoints by up to 3x.
- Deployed substantial server configuration changes with zero downtime using Kubernetes, Helm, and Pulumi.
- Technologies Used: Rust, Python, gRPCs, Protobufs, Networking, gVisor, Kubernetes, Helm, Pulumi.

Software Engineer Internship | Hudson River Trading May 2024 – Aug 2024

- Completed 8 projects in 11 weeks as a C++ Core Developer intern.
- Built an extensible, high-performance, multi-terabyte aggregate log viewer, saving hours of developers' time.
- Optimized the performance of a multithreaded searching tool, reducing CPU usage by more than 50%.
- Optimized our memory profiler, reducing the profiling overhead by more than 5x.
- Refactored a critical program to process data asynchronously, preventing it from intermittently hanging.

Software Engineer Internship | Codeium Jun 2023 – Aug 2023

- Led the initiative to quantize a large language model built with PyTorch C++, CUDA, CUBLAS, and Cutlass, achieving a ~2.3x speed increase for our company’s product in just 11 weeks.
- Researched, implemented, and evaluated various quantization methods (QAT, PTQ, int8, int4, NormalFloat, etc.), comparing performance and accuracy tradeoffs.
- Coordinated with a team of interns to write custom Cutlass kernels for quantized matrix multiplications, resulting in a ~2x speedup with minimal accuracy loss. Wrote extensive documentation for future maintainers.
- Identified performance bottlenecks with Nvidia Nsight Systems profiler, fused operations with custom CUDA kernels, optimized memory traffic, and integrated FlashAttention to further increase performance by ~1.15x.
- Collaborated with other interns to migrate our Jax training stack to use PyTorch, enhancing GPUs performance.
- Worked on a distributed MapReduce data processing pipeline with Go and Kubernetes.
- Technologies Used: PyTorch, Python, C++, CUDA, Cutlass, LLMs, Quantization, Jax, Go.

Co-Founder, Lead Developer | USACO Guide May 2020 – Dec 2022

- Designed, developed, and maintained an open-source competitive programming training website. 70,000+ users, 400,000+ monthly pageviews, 300+ contributors, 1,500+ GitHub stars, and 4,000+ Pull Requests.
- Created a real-time collaborative online IDE (75,000+ users) and serverless code execution system, running 6,000,000+ submissions at a cost of ~$0.06 per 1,000 executions.
- Founded a nonprofit organization with 50+ volunteers aiming to make competitive programming accessible to everyone. Organized classes, webinars, clubs, and more, impacting 50,000+ students.
- Technologies Used: React, Gatsby, AWS Lambda, Node.js, Tailwind CSS, Firebase, Typescript, MDX.