(718) 666-1912 | yuweis2@illinois.edu | linkedin.com/in/yuwei-sun-1136b0217/ | Champaign, IL

NYU High Performance Computing (RIT) Sep 2024 - May 2025

- Drove the migration from a physical HPC cluster to an on-premises cloud by automating a CI pipeline for deploying a reproducible environment on OpenShift Kubernetes, reducing onboarding time by over 40%.
- Deployed secure LLM inference infrastructure using Podman and Quay on Kubernetes with load-balancing and GPU scheduling; integrated vLLM to boost inference throughput by 67% compared to Ollama.
- Benchmarked HPC apps on ARM (GCP Axion, Nvidia Grace-Hopper) to evaluate price-performance efficiency.

NYU High Performance Computing (RIT) May 2024 - Sep 2024

- Delivered Reform v1.1.0 using Biopython and parallelized I/O optimization; added sequential processing, noval chromosome, and CI automation, boosting editing performance by 20%.
- Launched a browser-accessible interface (ReformWeb) for Reform using Flask, SQLite, and Redis, streamlining access to genome editing tools for 10+ research labs without local setup.

New York University Jan 2024 - May 2024

- Addressed AMD GPU underutilization by compiling and containerizing 5 GPU workflows (e.g., AlphaFold2, PyTorch) with AMD GPU support using Conda and Singularity, enabling rapid deployment on AMD nodes.
- Ported CUDA-based apps (e.g., DualSPHysics) to AMD's ROCm platform using HIPify; collaborated with the AMD HIPify team to upstream a patch supporting math_constants.h.

Aug 2025 - May 2027

Master of Computer Science

Sep 2021 - May 2025

Bachelor of Science in Computer Science, Minor in Mathematics *GPA: 3.78 (Dean's List)*

Feb 2025 - May 2025

- Led a 4-person team to build a serverless travel planner using AWS Lambda, S3, RDS, Cognito, and API Gateway to simplify multi-day trip planning.
- Integrated LLM with prompt engineering for trip recommendations and paired it with OCR-based ticket parsing.
- Architected an auto-scaling backend with ElastiCache, EventBridge, and SQS, reducing API response time by 45% and scaling to support 1000+ concurrent users in load tests.

Jun 2024 - Dec 2024

- Built an Ethernet-connected 6-node Raspberry Pi cluster with an AI accelerator and shared SSD storage via NFS.
- Integrated MPI and Kubernetes for distributed computing and enabled remote access via FRP and static IP.

May 2023 - Aug 2023

- Redesigned a monolithic order system into a modular DDD architecture to enhance scalability and maintainability.
- Scaled system throughput from 1,000 to 50,000 QPS by integrating Redis caching, RocketMQ messaging, and MySQL replication; enabled async order processing and TCC-based inventory management for high reliability.

- Uber Career Prep: Software Engineering Fellowship (Jan 2023)
- Student Cluster Competition (SCC23) – 6th Place Globally (Nov 2023)
- University Honors Scholar - Class of 2025

- Python, C++, Bash, SQL, CUDA, HIP, Java, Go, Flask, PyTorch, NumPy
- Linux, Docker, Podman, Kubernetes, SLURM, MPI, AWS, GCP, Git, Redis, CI/CD