

YUXIN(ASHLEY) YE

San Francisco, CA, USA yeyuxin9816@gmail.com (+1) 508-745-8907

INDUSTRY EXPERIENCES

Cerebras System - Member of Technical Staff II

Jan, 2025 - Now

- Deployed datacenter-scale Cerebras clusters for reliable inference infrastructure.
- Built interactive inference pipelines on custom wafer-scale ASIC hardware.
- Accelerated MoE models with HPC techniques under the weight streaming framework..

Lamini AI (Acquired by AMD) - High Performance Computing Engineer

May, 2024 - Jan, 2025

- Worked with Dr.Gregory Diamos: enabled RoCE RDMA networks over AMD Mi300 and Mi250 with Mellanox/Broadcom driver integrated with Kubernetes and Slurm nodes.

- Enabled Data Parallelism and Fully Sharded Data Parallelism (FSDP) with mpi4py using 100Gbps RoCE network.

- Deployed vLLM compatible pods with different models in NVIDIA and AMD machines.

Amazon Robotics - Software Development Engineer

July, 2022 - March 2024

- Added multiple features for event-driven virtual system simulator of automation sortation floor, including drive path planning, staging cell distribution, resource-based allocation algorithm, metric publisher...

EDUCATION

Master of Science

2020 - 2022

Computational Science and Engineering, GSAS, Harvard University, Cambridge, United States

GPA : 3.87

Bachelor of Science in Physics

2016 - 2020

School of the Gifted Young, University of Science and Technology of China, Hefei, China

GPA : 3.7

PROJECTS

Simulation of Soft Fish-like Swimmers (C++) | Research Assistant

May 2021 - November 2022

Github: <https://github.com/YYXLN/CUP2D>

Advisor: Professor Petros Koumoutsakos, SEAS, Harvard University

- Incorporated the elastic interaction between fluid and soft body using inverse map technique in a direct numerical simulators of flows based on the ACM Gordon Bell Prize-winning Cubism library with high-performance computing framework using Adaptive Mesh Refinement method.

- Parametrized spline movement by 7 neural knots and used reinforcement learning to train the swimming behaviour through muscle movement.

Gomoku Solver | MIT 6.877 Principles of Autonomy and Dec. Making

November 2021 - January 2022

- Implemented Upper Confidence Bound Monte Carlo Tree Search (UCT) for solving 9×9 Gomoku game.

- Used Neural Network as default policy in UCT .

Wasserstein Learning of Generative Models | MIT 6.838 Shape Analysis

April 2021 - May 2021

- Constructed the WGAN-GP from scratch and validate the tractable loss function of WGAN-GP. Offered a way to compute a Monge map.

- Presented several problems of WGAN-GP and give possible explanations assisted by comparison with the results from W2GAN.

- Presented experiments on high dimensional datasets including MNSIT and CIFAR10.

Realtime snow simulation and rendering (C++) | MIT 6.837 Computer Graphics

November 2020 - December 2020

- Used moving least square material point method to simulate the snow.

- Won a runner-up in MIT 6.837 (Computer Graphics) course project contest.

SKILLS

Programming: C/C++, Go, Python, Java, Matlab, Julia, Kotlin, Lisp, Haskell.

Software/ Framework: Spring, Pytorch, Jax, Triton, RoCm, OpenGL, CDK, Linux, Git, COMSOL, Latex, Lex, Spark, Map-Reduce, AWS, OpenACC, OpenMP, MPI, DVC, React, FastAPI.