TIANYUAN WU

+86 19953165605 ◇ Shanghai, China
wuty@shanghaitech.edu.cn ◇ linkedin.com/in/tianyuan-wu-578b4b1b0

EDUCATION

Master of Computer Science, ShanghaiTech University — Expected 2024
GPA: 3.84/4.00, Ranking: 5/340 (Top 2%)

Bachelor of Computer Science, ShanghaiTech University — 2017 - 2021
GPA: 3.8/4.00, Ranking: 2/209 (Top 1%)

SKILLS

Programming Languages	Familiar: Python, C, C++; Intermediate: Rust, JavaScript
Familiar Frameworks	PyTorch, OpenMP, MPI, OpenCV, OpenGL, Vulkan, CUDA, Vue.js, Gym
Interests	Machine Learning Systems, Reinforcement Learning, ML acceleration

EXPERIENCE

Research Intern — May 2022 - July 2023
Microsoft Research — *Beijing, China*

- Designed and implemented Catur, a reinforcement learning-based virtual NUMA placement system for Microsoft Azure. Catur was successfully integrated into Hyper-V and is currently being tested on production systems.
- Reduced the defective VM ratio of Azure general purpose clusters by 77.0% and tickets ratio by 75.2% using a novel and robust reinforcement learning method with continuous learning capabilities.
- Conducted research on continuous-time deep learning techniques, including neural ordinary differential equations (ODEs), control differential equations (CDEs), and state space models.
- Developed reinforcement learning algorithms for irregular-sampled states based on state space models.

Teaching Assistant — August 2021 - May 2022
New York University Shanghai — *Shanghai, China*

- Designed homework and discussions for Discrete Mathematics and Computer Programming courses.
- Developed and maintained an online grading system for programming course using Python and Vue.js.
- Communicated effectively with foreign students and professors from diverse backgrounds and cultures.

Research Intern — July 2020 - October 2020
Institute of Computing Technology, Chinese Academy of Sciences — *Beijing, China*

- Conducted research on memory prefetching and memory access pattern analysis techniques at the Center for Advanced Computer Systems, National State Key Lab of Computer Architecture.
- Developed a memory access pattern analyzer based on the ROSE compiler infrastructure that effectively captured and classified complex memory access patterns, enabling more efficient prefetching strategies to be developed.

PROJECTS

Catur: a reinforcement learning based virtual NUMA placement system of Azure.

- Used state-of-the-art reinforcement learning algorithms and continuous learning methodologies to develop a robust and efficient VM placement system.
- Catur was integrated into Hyper-V and tested on large-scale production systems, resulting in significant improvement in VM performance.
- Our paper on this project is currently under review for SOSP'23.

Portus: an efficient DNN checkpointing system with zero-copy RDMA.

- Developed a peer-to-peer RDMA-based datapath between GPUs and persistent memories, which enables zero-copy checkpointing of deep neural networks.
- Portus significantly improved checkpointing efficiency, achieving up to 8× faster checkpointing for distributed large model training (e.g., GPT) with its serialization-free index structure and zero-copy datapath.
- Our paper on this project is currently under review for NSDI'23.

Dash: a learning-based cluster scheduler for heterogeneous resources.

- Developed Dash, the scheduler for the HPC and AI cluster of ShanghaiTech University based on Slurm.
- Dash incorporated a lightweight neural network to predict job efficiencies on heterogeneous hardware by analyzing the CUDA kernels used. Based on these predictions, Dash provided efficient job scheduling policies with an effective reinforcement learning method, which optimized the average wait time and system efficiency.

Raster: a region-aware data management middleware for scientific computations.

- Developed Raster, a data management middleware for self-describing files such as HDF5 and netCDF. Raster is capable of efficiently handling queries for irregular-shaped regions among large-scale scientific datasets.
- Raster was deployed and tested on the Sunway TaihuLight supercomputer with the spatial-temporal datasets generated by the climatology application CESM. It showed a remarkable speedup of 5.6x on regional queries.

ACTIVITIES

- As a member of the ShanghaiTech GeekPie HPC team, participated in the Asia Supercomputing Competition (ASC'20) and improved the performance of BERT on a cluster for a CLOZE job.
- Also as a member of the ShanghaiTech GeekPie HPC team, participated in the Student Supercomputing Competition (SCC'20) and worked on optimizing the computational biology software, GROMACS. Published a critique on TPDS'20: Reproducibility: Performance Evaluation of MemXCT on Azure CycleCloud Platform
- Currently serving as a student advisor for ShanghaiTech University for the Asia Supercomputing Competition (ASC'23).

AWARDS

- Outstanding Teaching Assistant Award, ShanghaiTech University, 2019.
- Second-class Scholarship, ShanghaiTech University, 2019.
- Outstanding Teaching Assistant Award, ShanghaiTech University, 2020.
- Outstanding Undergraduate Thesis, ShanghaiTech University, 2021.