Shaoyu Wang

Luoyu Road 1037, Wuhan, Hubei 430074, China
wsyy0619@gmail.com • (+86) 136-9813-2563

EDUCATION

Huazhong University of Science and Technology (HUST) Wuhan, China
- B.E. in Computer Science and Technology, GPA: 3.72 / 4.0 Sep 2020 – Jun 2024 (Expected)

RESEARCH INTEREST

My current focus is improving **AI system performance** across the whole computing stack from large scale training and serving to hardware/software co-design. I am also enthusiastic in **computer architecture**, specifically general-purpose GPU (GPGPU) architecture and Domain-Specific Accelerators (DSA).

PUBLICATIONS

Sequence Batching with Prediction: A Speculation-based LLM Batch Serving System
Shaoyu Wang, Ziming Liu, Yang You
In progress for ICML2024

Visual Exploratory Analysis for Designing Large-Scale Network-on-Chip Architectures: A Domain Expert-Led Design Study
Shaoyu Wang*, Hang Yan*, Katherine E. Isaacs, Yifan Sun
IEEE Transactions on Visualization and Computer Graphics (TVCG)

RESEARCH EXPERIENCE

HPC-AI Lab at National University of Singapore (NUS) Jul 2023 – Present
Research Intern, Advised by **Prof. Yang You**
Speculation-based sequence batching for **Large Language Model (LLM) serving**
- Created an asynchronous serving system that incorporates an optimized dual-stage pipeline of prediction and serving processes, improving system throughput.
- Developed a dynamic programming based sequence batching algorithm, utilizing pre-serving profiling to improve efficiency and performance of batch computing.
- Established a mechanism to detect and reprocess sequences with prediction failure, ensuring reliable and consistent serving outcomes.

Scalable Architecture Lab at College of William & Mary Apr 2022 – Mar 2023
Research Intern, Advised by **Prof. Yifan Sun** and **Prof. Katherine E. Issacs**
Wafer-Scale **GPU architecture** performance analysis and optimization
- Proposed a load-balanced Network-on-Chip (NoC) routing policy, improving NoC bandwidth utilization.
- Introduced a metrics collector within the MGPUSim simulator to gather network traffic data.

Visualization of mesh **Network-on-Chip (NoC)** traffic
- Integrated with an architecture-level profiling tool to gain insights of performance bottlenecks.
- Developed a systematic visualization solution combining temporal and spatial views for effective hotspot identification in the network.

WORK EXPERIENCE

NVIDIA Shanghai, China
Compute Architect Intern at HPC Architecture Performance Team Mar 2023 – Jul 2023
Enhanced the internal-used simulation tool of **GPGPU architecture** prototyping and performance analysis
- Optimized the simulation process, significantly reducing the time required from more than a day to just several minutes for workloads involving over 400 traces.
- Expanded the simulator's capabilities by adding support for the cutting-edge NVIDIA Hopper architecture, enabling performance analysis of latest GPU cards
- Introduced new evaluation models for hardware load balance and Streaming Multiprocessors (SM) utilization, boosting the precision and reliability of simulation.

ByteDance (TikTok) Hangzhou, China
Software Engineering Intern at Cloud Computing Infrastracture Team Jun 2022 – Sep 2022
Enhanced KubeBrain, a high-performance **distributed data management system** for Kubernetes
- Developed an backup and recovery module in KubeBrain, bolstering its reliability and resilience.
- Integrated a hot cache mechanism to efficiently handle time-consuming queries, reducing the request latency by 12x.

OPEN SOURCE CONTRIBUTION

OIWiki

Contribute to OIWiKi, the online wiki for competitive programming comptetions

- Developed a plugin to support the compilation of Markdown tab syntax, enabling the OIWiki community to efficiently process documents featuring content tabs
- Enhanced the export tool-chain for converting into LaTeX and Typst formats, extending export capabilities of OIWiki materials

OpenEuler

Contribute to iSulad, the open source **container manager** written in C++

- Extended the communication protocol of iSulad, facilitating seamless interaction between the command-line interface (CLI) and the backend server
- Implemented progress monitoring for container image pulling, enhancing its ability to track and display real-time progress

STUDENT CLUSTER COMPETITIONS

ISC Student Cluster Competiton 2023, virtual track Finalist

Build up and lead a student group of 6, advised by **Prof. Xuanhua Shi** and **Prof. Yao Wan**

- Deployed and profiled 3 scientific computing application (Quantam Expresso, FluTAS, POT3D) using several computing nodes in supercomputer centers.
- Managed to build CPU-GPU hybrid programs and orchestrated them over 4 computing nodes using MPI and InifiniBand (IB).

ASC Student Supercomputer Challenge 2022 Second Prize

Being one of a student group of 5, advised by **Prof. Xuanhua Shi**

- Pre-trained a 4.7B language model Yuan-1.0 on the given 100GB dataset using Megatron, orchestrating 3D parallelism for distributed training over 8 nodes
- Optimized DeepMD training process, achieved 4x speedup on CPU and 2x speedup on GPU

3rd "AMD Cup" International Parallel Computing Competition Silver Medal

- Accelerated a high-dimension point cloud feature extractiona algorithm, achieving 1200x speedup
- Accelerated the SOTA graph spectral sparsification algorithm feGRASS, achieving 40x speedup

9th "Intel Cup" Parallel Application Challenge Bronze Medal

- Optimized an application for genome sequence analysis, achieving 13x performance improvement

SKILLS

Programming Languages: C/C++, Python (pytorch), Golang, JavaScript/TypeScript, Verilog
Computer Architecture: Software Simulation, High level synthesis (HLS)
High Performance Computing: OpenMP, MPI, SSE/AVX, CUDA, Code profiling
Large Language Model: Parameter efficient fine-tune (PEFT), Data/Pipeline/Model Parallelism

HONORS & AWARDS

2023 ISC Student Cluster Competiton May 2023
- Finalist of the online track

2022 ASC Student Supercomputer Challenge April 2022
- Second Prize in the preliminary contest

The 3rd "AMD Cup" International Parallel Computing Competition Silver Medal, China Nov 2022
- Rank 3rd in the final, competed with about 120 teams in China

The 9th "Intel Cup" Parallel Application Challenge Bronze Medal, China Oct 2021
- Rank 5th in the final, competed with about 120 teams in China

Huawei Intelligent Base Scholarship May 2022
- Top 0.5%, for outstanding innovation and strong practical skills of CS students

Scientific and Technology Innovation Scholarship, HUST Dec 2021
- Top 3.0%, award for scientific research and innovation of undergraduates