Ao Shen

shen634@purdue.edu | https://aoshen524.github.io

Education

Purdue University (Transfer) West Lafayette, IN

Bachelor of Science in Computer and Information Technology; Minor in Mathematics (GPA: 3.69) *Aug 2022 – May 2025*

Research Interests

System & Architecture for Emerging Workloads, HPC for Science.

Research Experience

Full-Time Research Assistant Shanghai Qi Zhi Institute

Advisor: Mingyu Gao, Assistant Professor at IIIS, Tsinghua University *May 2023 – Aug 2024*

LLM Inference System Research and Development: FastSwitch

- First-author paper published: https://aoshen524.github.io/files/mlsys.pdf.
- Under guidance, developed an inference system based on vLLM, designed for high-frequency preemption and multi-turn conversations.
- Asynchronous Operator Dispatch with Multi-threading, Multi-stream, and Graph Integration: Overcame Python's Global Interpreter Lock (GIL) limitations by enabling asynchronous custom operator dispatch through multi-threading in C, allowing full overlap of I/O and compute operations.
- Task Handling and Conflict Resolution: Enhanced system efficiency by separating context switch tasks from regular storage tasks. Developed a cache reuse strategy that significantly reduced context switch overhead.
- Advanced System-Level Memory Management: Observing bottlenecks caused by dispatch-induced swapping, integrated the existing paged-attention strategy with an I/O-aware memory allocation system for KV cache. This improved memory continuity for tasks and combined the benefits of dynamic memory pools and buddy allocators, enhancing I/O performance without introducing additional overhead.
- Impact: Explored the integration of our work into service scenarios with rapidly shifting priorities to ensure fairness in meeting service-level objectives without compromising performance. FastSwitch achieved a speedup of 1.4×–11.2× in TTFT & TBT tail latencies across percentiles, with no more than 1% additional call stack overhead.

Neural Architecture Search Research and Development: Canvas

- Third-author paper published: https://arxiv.org/pdf/2304.07741.
- Reproduced multiple NAS frameworks, including DARTS and EoiNAS, and integrated them into an online kernel profiling and selection system. This enabled efficient evaluation and optimization of kernels within the NAS workflow.
- Efficiently selected kernels by treating them as NAS edges, avoiding full training for each kernel to obtain its accuracy on downstream tasks. This approach identified a set of high-precision, high-performance kernels from a pool of over 100,000 candidates.
- Impact: Achieved speedups of up to 2.3× in experiments, while maintaining accuracy.

Research Assistant Purdue University

Advisor: Baijian Yang, Professor, Purdue University *Jan 2023 – Mar 2023*

- Implemented and deployed the buffer overflow attack lab and ROP attack lab with containerization technologies such as Docker & K8s, allowing users to complete the learning of common network attacks and defenses without the need for significant resources.

Asia Student Supercomputer Challenge, First Prize Nov 2021 – Mar 2022

- Set up SSH communication between two nodes and created NFS shared storage across the nodes.
- Tested the computing performance of our equipment using HPL and HPCG, achieving 95% of the official benchmark performance by controlling key variables.
- Built a software development environment for the DeepMD-Kit, a C++-based software used in molecular dynamics simulations.
- Applied parallel optimization techniques based on underlying source code principles and relevant disciplinary knowledge.
- Utilized CPU parallel programming optimizations such as AVX vectorization and OpenMP. By employing OpenMP techniques, achieved a 25% speedup, and further improved performance by 50% through code structure optimization and removal of redundant code.

Skills

Tools: Python, C++, CUDA, PyTorch, Nsight System, Nsight Compute, Hugging Face-related libraries.

Languages: Mandarin (Native), Cantonese (Very Fluent), English (Very Fluent). Served as an interpreter at the world's largest trade fair.

Honors and Awards

Dean's List: 2022, 2023

Major contributor and admin of CUDA group chat, AI inference and deployment group chat, architecture group chat.