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 and Architecture for Emerging Workloads, Edge Computing.

Publications

- **FastSwitch: Optimizing Context Switching Efficiency in Fairness-aware Large Language Model Serving.**
  Ao Shen, Zhiyao Li, Mingyu Gao.
  arXiv preprint: arXiv:2411.18424.
  Contribution: During my internship at Shanghai Qi Zhi Institute, under the guidance of my advisor, I explored the research problem, implemented the methods, conducted the experiments, and wrote the paper.

- **Canvas: End-to-End Kernel Architecture Search in Neural Networks.**
  Chenggang Zhao, Genghan Zhang, **Ao Shen**, Mingyu Gao.
  arXiv preprint: arXiv:2304.07741.
  Contribution: During my internship at Shanghai Qi Zhi Institute, under the guidance of my advisor, I explored the research problem, implemented the methods, and conducted the experiments.

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

- 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 multithreading in C++, allowing full overlap of I/O and inference operations.
- Task Handling and Conflict Resolution: Enhanced system efficiency by separating context switching tasks from regular storage tasks. Developed a cache reuse strategy that significantly reduced context switch overhead.
- Advanced System-Level Memory Management: Observed bottleneck caused by dispatch-induced swapping. Integrated the existing paged attention strategy with an I/O-aware memory allocation system for KV cache. Improved KV cache continuity and combined the benefits of dynamic memory pools and buddy allocators. Enhanced I/O performance with little overhead.
- Impact: Explored the integration of our work into service scenarios with rapidly shifting priorities to ensure fairness in meeting service-level objective without compromising performance. Achieved a speedup of 1.4–11.2× in TTFT and TBT tail latencies across different models and scheduling policies, with no more than 1% additional call stack overhead.

Neural Architecture Search Research and Development: Canvas

- Reproduced multiple NAS frameworks, including DARTS and EoiNAS, and integrated them into an online kernel profiling and selection system. 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. 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.

Project Experience

Computer Architecture Project Purdue University
Advisor: Kazem Taram, Assistant Professor at Purdue University *Sep 2024 – Nov 2024*

- Designed and implemented branch prediction algorithms (Bimodal and Gshare). Evaluated predictors on real program traces, reducing misprediction rates and analyzing trade-offs.
- Developed a cache simulator to compare replacement policies (LRU, Random, Tree-PLRU) and implemented a striding prefetcher, reducing cache miss rates. Optimized CPU kernel memory locality, cutting miss costs by 50%, and documented performance trends.
- Implemented Flush+Reload and Prime+Probe attacks to exploit cache timing vulnerabilities, recovering cryptographic keys. Built a covert-channel mechanism using cache timing, achieving reliable data transmission and evaluating bandwidth under noise.

Network Security Project Purdue University
Advisor: Baijian Yang, Professor at Purdue University *Jan 2023 – Mar 2023*

- NSF-funded project: Collaborative Research - CHEESE: Cyber Human Ecosystem of Engaged Security Education. https://www.cheesehub.org/en/latest/cheesehub.html.
- Implemented the buffer overflow attack lab and Return-Oriented Programming attack lab.
- Deployed the labs using containerization frameworks such as Docker and K8s, enabling resource-efficient execution and scalability for learners without requiring significant local resources.

Autonomous Robot Project Purdue University

Advisor: Byungcheol Min, Associate Professor at Purdue University *Jan 2023 – Mar 2023*

- Designed and programmed a simulated autonomous mobile robot using C++ to navigate unknown environments, measure temperature and light levels with thermistor and photocell sensors, and avoid obstacles using an ultrasonic sensor.
- Integrated a servo motor and DC motor to control the robot's direction and movement, leveraging ultrasonic sensor inputs to adjust the rotational speed and steering angle for obstacle avoidance.

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 cluster using HPL and HPCG, achieving 95% of the official benchmark performance by controlling key variables.
- Configured and deployed a fully integrated development environment for DeepMD-Kit, a C++ framework for molecular dynamics simulations, streamlining the workflow for large-scale simulations
- Applied parallel optimization techniques based on underlying source code principles and relevant disciplinary knowledge.
- Leveraged CPU parallel programming optimizations, including AVX vectorization and OpenMP directives. Achieved a 25% performance improvement through these optimization. Further enhanced system throughput by 10% via code refactoring and elimination of computational redundancies.

Skills

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

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

Honors and Community Contribution

Dean's List: 2022, 2023

Academic Excellence Award: 2022

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