Zhuofu Chen

aetiurf@gmail.com · chenzhuofu · https://chenzhuofu.github.io/ ·

Education

Tongji University Shanghai, China

B.S. in Computer Science and Technology (Elite Class) Sept. 2021 - Present

- GPA: 4.87/5.00 Ranking: 3/20

Research Interests

I have broad interests in building system infrastructures to systematically bring better *performance* and *usability* to *real-world applications*. Specifically, I often contemplate how to *redesign next-generation datacenter/cloud operating systems* to bridge the gap between existing hardware and emerging needs of software, and to serve numerous applications such as AI inference/training and cloud computing.

Experience

Catalyst Group, Carnegie Mellon University

Research Intern advised by Zhihao Jia May. 2024 - Present

- co-designed a request level scheduler that prioritizes requests considering their proximity to SLO violations, achieving 20-30% higher SLO attainment.
- Reimplemented the main Transformer operators such as attention and allreduce to get about 10x latency improvement for kernel execution in large batch decoding and achieved state-of-the-art serving performance.
- One paper is pending review.
- Actively involved in other efficient AI applications projects such as TidalDecode.

Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University

Research Intern advised by Xingda Wei and Haibo Chen Jul. 2023 - Apr. 2024

- Developed an optimal GPU disaggregation system for transparently and efficiently serving AI applications.
- Created a theoretical model characterizing the overhead of disaggregating applications, achieving an error margin within 10%.
- Studied the lower bound of datacenter network for supporting GPU disaggregation.
- One paper is pending review.

Key Laboratory of Embedded System and Service Computing, Tongji University

Research Intern advised by Zhijun Ding Nov. 2022 - Oct. 2023

- Implemented a WebAssembly-based runtime with an OCI shim to bridge orchestration tools and runtime.
- Invented a dynamic-import mechanism enabling multiple memory-sharing for WebAssembly modules.
- Supported shared memory and TCP/IP communication across WebAssembly modules.

Selected Projects

XPURemoting, a performant GPU disaggregation system

- Hijack the CUDA driver API to transparently redirect GPU calls.
- Propose a concise and elegant network abstraction for GPU disaggregation.
- Introduce an easy-to-use perf tool to model the overhead of disaggregation for arbitrary applications.
- 4k LOC in C++ (v1), 15k LOC in Rust (v2).

FlexFlow SpecScheduler, a SLO-aware scheduler for serving large model inference

- Introduce a request-level scheduler towards optimal SLO attainment and goodput.
- Deliver state-of-the-art kernel performance for large batch speculative decoding.
- Achieve 20-30% higher SLO attainment compared to existing serving systems.
- 10k LOC in C++/CUDA.

Selected Awards

Tongji University Lu Hao Scholarship	2024
National 1st Prize (0.55%) in Contemporary Undergraduate Mathematical Contest in Modeling	2023
Regional 1st Prize in Contemporary Undergraduate Mathematical Contest in Modeling	2023
China National Scholarship (top 0.2%)	2022
Regional 2nd Prize in Contemporary Undergraduate Mathematical Contest in Modeling	2022
Bronze Medal of National Olympiad in Informatics (NOI)	2020
1st Prize of National Olympiad in Informatics in Provinces	2019