

# 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, scalability* and *usability* to *real-world applications*. Specifically, I often contemplate how to *redesign next-generation datacenter systems* to bridge the gap between evolving hardware and emerging needs of software, and to serve data-intensive applications such as *AI inference and training*.

## EXPERIENCE

---

### Catalyst Group, Carnegie Mellon University

*Research Intern* advised by [Zhihao Jia](#)

May. 2024 - Present

- Optimized the LLM serving system for large batch decoding, achieving 7x latency improvement.
- Designed a request level scheduler considering different SLO requirements across requests, achieving up to 73% higher SLO attainment.
- Implemented a novel position-persistent mechanism to reduce the overhead of sparse attention, demonstrating 20% improvement over the state-of-the-art.
- Two papers are pending review.

### Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University

*Research Intern* advised by [Xingda Wei](#) and [Rong Chen](#)

Jul. 2023 - Apr. 2024

- Developed an optimal GPU disaggregation system for transparently and efficiently serving AI applications with no performance degradation compared to local execution.
- Created a theoretical model characterizing the overhead of disaggregating applications, achieving an error margin within 10%.
- Built a profiling-based tool that allows users to derive network requirements for any AI applications.
- 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

- Proposed a dynamic import mechanism to enhance WebAssembly-based runtime, introducing a shareable resource allocation paradigm.
- Implemented a state transfer method through shared memory, improving serialization into a storage layer.
- Implemented an OCI shim to bridge orchestration tools and runtime.

## PUBLICATIONS

---

(\* indicates equal contribution)

- [1] Tianxia Wang\*, **Zhuofu Chen\***, Xingda Wei, Jinyu Gu, Rong Chen, Haibo Chen. [Characterizing Network Requirements for GPU API Remoting in AI Applications](#), *Under review*.
- [2] Zikun Li\*, **Zhuofu Chen\***, Remi Delacourt, Gabriele Oliaro, Zeyu Wang, Qinghan Chen, Shuhuai Lin, April Yang, Zhihao Zhang, Zhuoming Chen, Sean Lai, Xupeng Miao, Zhihao Jia. [AdaServe: SLO-Customized LLM Serving with Fine-Grained Speculative Decoding](#), *Under review*.
- [3] Lijie Yang\*, Zhihao Zhang\*, **Zhuofu Chen**, Zikun Li, Zhihao Jia. [TidalDecode: Fast and Accurate LLM Decoding with Position Persistent Sparse Attention](#), *Under review*.

## SELECTED PROJECTS

---

### XPURemoting

- A GPU disaggregation system redirecting CUDA API calls to remote devices with near-native performance.
- An easy-to-use perf tool to model the overhead of disaggregation for arbitrary applications.
- 15k lines of code in Rust.

### FlexFlow SpecScheduler

- A performant LLM serving system for large batch decoding, especially for speculative decoding.
- An request-level scheduler aiming to maximize SLO attainment across diverse tasks.
- 10k lines of code in C++/CUDA.

## SELECTED AWARDS

---

SOSP Student Scholarship	2024
Tongji University Lu Hao Scholarship	2024
National 1 <sup>st</sup> Prize (0.55%) in Contemporary Undergraduate Mathematical Contest in Modeling	2023
Regional 1 <sup>st</sup> Prize in Contemporary Undergraduate Mathematical Contest in Modeling	2023
China National Scholarship (top 0.2%)	2022
Regional 2 <sup>nd</sup> Prize in Contemporary Undergraduate Mathematical Contest in Modeling	2022
1 <sup>st</sup> Prize of National Olympiad in Informatics in Provinces	2019