Boyu Tian (田博宇)

E-mail: tby20@mails.tsinghua.edu.cn ***Homepage:*** https://criust.github.io ***Telephone/Wechat:*** +86-15389218086

EDUCATION

Tsinghua University (清华大学) *Sep. 2020 - Present*
Institute for Interdisciplinary Information Sciences (交叉信息研究院)
Ph.D. student in Computer Science, advised by Prof. Mingyu Gao (高鸣宇).

Shanghai Jiao Tong University (上海交通大学) *Sep. 2016 - Jun. 2020*
ACM Honors Class, Zhiyuan College (ACM 班)
B.Eng. in Computer Science, advised by Prof. Chao Li (李超).

RESEARCH INTERESTS

My research interests mainly lie in efficient memory architectures and scalable data processing, with a focus on memory-centric designs like near-data processing, memory pooling, and memory disaggregation. Additionally, I explore memory system optimizations for key workloads like graph processing and large language models.

PUBLICATIONS

Boyu Tian, Yiwei Li, Li Jiang, Shuangyu Cai, and Mingyu Gao. NDPBridge: Enabling Cross-Bank Coordination in Near-DRAM-Bank Processing Architectures. In ***ISCA***, 2024. (**CCF-A**)

Boyu Tian, Qihang Chen, and Mingyu Gao. ABNDP: Co-optimizing Data Access and Load Balance in Near-Data Processing. In ***ASPLOS***, 2023. (**CCF-A**).

Shuangyu Cai, **Boyu Tian**, Huanchen Zhang, and Mingyu Gao. PimPam: Efficient Graph Pattern Matching on Real Processing-in-Memory Hardware. In ***SIGMOD***, 2024. (**CCF-A**).

Qihang Chen, **Boyu Tian**, and Mingyu Gao. FINGERS: Exploiting Fine-Grained Parallelism in Graph Mining Accelerators. In ***ASPLOS***, 2022. (**CCF-A**).

Bohan Zhao, Xiang Li, **Boyu Tian**, Zhiyu Mei, and Wenfei Wu. DHS: Adaptive Memory Layout Organization of Sketch Slots for Fast and Accurate Data Stream Processing. In ***KDD***, 2021. (**CCF-A**)

RESEARCH EXPERIENCES

IDEAL Lab, IIIS, Tsinghua University *Sep. 2020 - Present*
Research Assistant, advised by Prof. Mingyu Gao *Beijing, China*

- We focused on alleviating the memory access bottleneck for data-intensive applications. I paid special attention to architectures that follow the Near-Data Processing paradigm. I aim at providing system support and data communication optimization for NDP systems with various hardware technologies, including 3D-stacked-memory-based NDP *(ASPLOS' 23)* and DRAM-bank-based NDP *ISCA' 24*.
- *(Ongoing)* I am currently working on alleviating the memory bottleneck of large language model inference with NDP hardware and data offloading .
- *(Ongoing)* I am currently working on architecting a high-performance rack-scale CXL-based memory pool that is scalable and applicable to heterogeneous devices and various applications.

SAIL Lab, Shanghai Jiao Tong University *Jul. 2018 - Jun. 2020*
Research Intern, advised by Prof. Chao Li *Shanghai, China*

- We explored the idea of approximate graph computing. I developed a system to control approximation level of graph algorithms according to user-defined QoS requirements.
- I proposed a graph abstraction for cloud resources and inter-dependent microservices, along with a microservice deployment scheme using sub-graph matching and a runtime resource adjustment.

CEI Lab, Duke University *Jul. 2019 - Sep. 2019*
Research Intern, advised by Prof. Yiran Chen *North Carolina, U.S.*

- I explored the idea of accelerating graph processing using ReRAM-based Processing-in-Memory paradigm.

INDUSTRY EXPERIENCES

Zhipu AI (智谱 AI) *Mar. 2024 - Present*
Research intern in AI Academy. Mentor: Dr. Guanyu Feng *Beijing, China*

- We focused on large language model inference acceleration.

Alibaba DAMO Academy (阿里巴巴达摩院) *Jun. 2023 - Jan. 2024*
Research intern in Computing Technology Lab. Mentor: Dr. Dimin Niu *Beijing, China*

- We focused on the design and development of memory pooling based on the CXL technology. Our work is presently being submitted to the industry track of leading conferences of computer architecture.
- I was in charge of a research project focusing on the design of a rack-level CXL-based memory pool that is scalable and applicable to multiple heterogeneous computing devices.

Turing Department, Huawei Hisilicon (华为海思图灵架构与设计部) *Oct. 2019 - Dec. 2019*
Research Intern, supervised by Dr. Heng Liao and Dr. Lin Li *Shanghai, China*

- I developed algorithms for 3D view synthesis from sparse input images. I modified the rendering path generation of existing synthesis systems to adapt it for light field rendering in the 3D scenario.

HONORS AND AWARDS

清华大学综合优秀奖学金	2021, 2022, 2023
ASPLOS 2023 Student Travel Award	2023
唐立新奖学金	2018-2020
上海交通大学致远杰出领袖奖学金	2017
上海交通大学致远荣誉奖学金	2016-2019

TEACHING

Teaching Assistant *20470084 Computer Architecture*
Spring 2021, Spring 2022 *Tsinghua University*

- I worked as the teaching assistant of Computer Architecture taught by Prof. Mingyu Gao, targeting undergraduate students in Yao Class and Artificial Intelligence Class in IIIS. I designed and developed the course project, which is a computer architecture simulator for RISC-V.

Teaching Assistant *C++ Programming*

Fall 2017 *Shanghai Jiao Tong University*

- I worked as the teaching assistant of C++ Programming taught by Prof. Huiyu Weng for students in ACM Class. I designed exam questions and algorithmic programming exercises.

TECHNICAL SKILLS

Programming Languages	C, C++, Python, Verilog, Java, Rust, Go
Hardware Simulation/Analysis	ZSim, Intel Pin, CACTI, Ripes