Shulai Zhang

No.800 Dongchuan Road, Shanghai, China | www.shulai.org | +86 189-0986-6566 | zslzsl1998@sjtu.edu.cn
Research interests: AI systems, GPU resource management/scheduling/compiling

EDUCATION

Shanghai Jiao Tong University — Shanghai, China
PhD candidate, Computer Science – GPA 3.78/4.0 — Sept. 2020 – Now
Shanghai Jiao Tong University — Shanghai, China
B.E., Information Engineering (AI oriented) – GPA 90.42/100 — Sept. 2016 – Jul. 2020

PAPERS

Shulai Zhang, et al., "Krypton: Enable Performance-Aware GPU Sharing with Functional Integrity and Isolation", In-submission

Shulai Zhang, Quan Chen, Weihao Cui, Han Zhao, Chunyu Xue, Zhen Zheng, Wei Lin, Minyi Guo, "Improving GPU Sharing Performance through Adaptive Bubbleless Spatial-Temporal Sharing", EuroSys 2025

Binghao Chen, Han Zhao, Weihao Cui, Yifu He, **Shulai Zhang**, Quan Chen, Zijun Li, Minyi Guo, "Maximizing the Utilization of GPUs Used by Cloud Gaming through Adaptive Co-location with Combo", SoCC 2023

Shulai Zhang, Weihao Cui, Quan Chen, Zhengnian Zhang, Yue Guan, Jingwen Leng, Chao Li, Minyi Guo, PAME: precision-aware multi-exit DNN serving for reducing latencies of batched inferences, ICS 2022

Shulai Zhang, Zirui Li, Quan Chen, Wenli Zheng, Jingwen Leng, Minyi Guo, "Dubhe: Towards data unbiasedness with homomorphic encryption in federated learning client selection", ICPP 2021

Shulai Zhang, Xiaoli Ma, "A General Difficulty Control Algorithm for Proof-of-Work Based Blockchains", ICASSP 2020

Kaiwen Zheng, **Shulai Zhang**, Xiaoli Ma, "Difficulty Prediction for Proof-of-Work Blockchains", SPAWC 2020

Shulai Zhang, Meixia Tao, and Zhiyong Chen, "Exploiting Caching and Prediction to Promote User Experience for a Real-time Wireless VR Service", GLOBECOM 2019

PROJECTS

Performance isolation tools for multi-tenant NPU — Project Leader
Institute of Computing Technology, Chinese Academy of Sciences, May 2024 – Now

Kernel-space virtualization for GPU sharing — Project Leader
Lenovo, Nov 2023 – July 2024

Resource management and compilation co-design to optimize AI model performance — Project Leader
Alibaba Group, May 2023 – May 2024

EXPERIENCES

ByteDance (Seed Foundation) – ML System Research Intern — July 2024 - Now
Agora (Real-Time Communications Group) – Research Intern — June 2020 – Oct. 2020
Georgia Institute of Technology – Research Intern — July 2019 – May 2020

SKILLS

Programming Languages: C/C++, CUDA, Python
Systems: GPU system stack, AI frameworks