

Shulai Zhang

No.800 Dongchuan Road, Shanghai, China | +86 189-0986-6566 | zslzsl1998@sjtu.edu.cn

Research interests: GPU/NPU resource management, ML systems

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
<i>Institute of Computing Technology, Chinese Academy of Sciences</i> , May 2024 – Now	
Kernel-space virtualization for GPU sharing	Project Leader
<i>Lenovo</i> , Nov 2023 – Now	
Resource management and compilation co-design to optimize AI model performance	Project Leader
<i>Alibaba Group</i> , May 2023 – May 2024	
Signal processing in Proof-of-Work blockchains	Research Intern
<i>Georgia Institute of Technology</i> , July 2019 – May 2020	

SKILLS

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

TEACHING

TA of Advanced Computer Architecture (CS7331)	2023
TA of AI Computing Systems (CS7345)	2021, 2022