TIANLE ZHONG

+01 434-833-0770 | zhluosuu@outlook.com | luosuu.github.io

EDUCATION

University of Virginia (UVA) **Charlottesville, USA**
PhD in Computer Science, Dept. Computer Science, School of Applied Science and Engineering *Sept 2022 to Now*
- Advisor: Prof. Geoffrey Fox

University of Electro-Communications (UEC) **Tokyo, Japan**
Exchange Student, School of Informatics *Oct 2020 to Aug 2021*

University of Electronics Science and Technology of China (UESTC) **Chengdu, China**
Bachelor of Computer Science and Applied Mathematics, School of Computer Science and Engineering *Sept 2018 to July 2022*

RESEARCH EXPERIENCE

AML Machine Learning System Group, ByteDance **San Jose, USA**
Research Interests: Large-scale Large Language Model (LLM) Training & Inference System *July 2025 to May 2026*
- Mentor: Xiao Yu (Head of Engineering)

Networking Research Group, Microsoft Research Asia (MSRA) **Beijing, China**
Research Interests: Large-scale Distributed Machine Learning System & Algorithms *Oct 2021 to May 2022*
- Mentor: Lei Qu (Senior Research Engineer)

UEC Haneda Sound Media Lab **Tokyo, Japan**
Research Interests: Speech Processing, Machine Learning, Learning Representation, Acoustics *Nov 2020 to Sept 2021*
- Advisor: Prof. Yoichi Haneda

SELECTED PUBLICATIONS

Youmu: Efficient Columnar Data Pipeline for LLM Training
Accepted by The Eighth Annual Conference on Machine Learning and Systems (MLSys 2025)
***Tianle Zhong**, Jiechen Zhao, Qiang Su, Geoffrey Fox*
- TL;DR: A data pipeline for LLM training on Parquet data to reduce storage cost and memory footprint for loading LLM datasets.

Proton: Towards Multi-level, Adaptive Profiling for Triton
Accepted by The IEEE/ACM International Symposium on Code Generation and Optimization (CGO 2026)
*Keren Zhou, **Tianle Zhong**, Hao Wu, Jihyeong Lee, Yue Guan, Yufei Ding, Corbin Robeck, Jeff Niu, Phil Tillet*
- TL;DR: Flexible and very low-overhead profiling for Triton kernels, built in Triton compiler infrastructure.

Reimu: Optimizing Data I/O for LLM Datasets on Remote Storage
Accepted by Cloud Intelligence/AIOps 2024 Workshop (Co-located with ASPLOS 2024)
***Tianle Zhong**, Jiechen Zhao, Xindi Guo, Qiang Su, Geoffrey Fox*
- TL;DR: A LLM training data pipeline designed for block-based storage with awareness of I/O efficiency.

Spherical Convolutional Recurrent Neural Network For Real-time Robust Sound Source Tracking
Accepted by 2022 IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP 2022)
***Tianle Zhong**, Israel Mendoza Velazquez, Ren Yi, Hector Manuel Perez Meana, Yoichi Haneda*
- TL;DR: a Laplacian graph-based spherical convolution to learn spatial stereoscopic features with SO(3) equivariance

NOTABLE PROJECTS

VeOmni: Scaling Any Modality Model Training with Model-Centric Distributed Recipe Zoo **ByteDance Seed**
Open Source Research Project, widely deployed inside ByteDance to empower multimodal MoE training. *2022.09 - Now*
- Implement an easy-to-use and efficient EP+FSDP2 mechanism for MoE model training at scale.
- Contribute and serve as an active maintainer.

PerfSim: Distributed Machine Learning System Performance Simulator **Microsoft Research Asia**
A research project at Networking Research Group, Microsoft Research Asia *2021.10 - 2022.05*
- Based on a graph-based computation operator flow profiler
- Extend the single node performance simulation to multi-node cases by a NCCL performance predictor

SKILLS, AWARDS & OTHERS

- **Programming:** Python (PyTorch), Rust, C & C++, Matlab, Java
- **Awards:** 2020 Most Valuable Member of Microsoft Student Club (issued by MSRA); Excellent Member of 2020 Tencent Cloud Development Summer Camp (10 of 300); UETSC College Pacesetter Scholarship; UEC Outstanding Student Certificate
- **Languages:** English: IELTS 7.0, GRE 321+3.5, have written 4 conference papers; Mandarin: native; Japanese: conversational.
- **Contests:** 2018 & 2019 UESTC Mathematical Modeling Contest (Second Award)