

Tiannuo Yang

Email

tnyang2000@gmail.com

Website

<https://tiannuo-yang.github.io>

Research Interests

Cloud Computing, Datacenter

ML4System, System4ML

Automated Performance Tuning

As Moore's Law fades, my work aims to fully exploit hardware resources, automate the operation of complex systems, and enhance the performance and efficiency of System4ML (e.g., LLM and RAG).

EDUCATION

Nankai University, China

Aug. 2022 ~ Present

Master in Computer Science (Exempted from Graduate Entrance Exam)

– Advisor: Prof. Yusen Li

University of Science and Technology Beijing, China

Aug. 2018 ~ June 2022

Bachelor in Information Management and Information System

– Cumulative GPA: **91.2**/100 Major GPA: **92.8**/100

Southern Taiwan University of Science and Technology, China

Sep. 2019 ~ Jan. 2020

Major in Information Management and Information System

– Cumulative GPA: **4.3**/4.3

Language: TOEFL iBT 104 (R27, L25, S22, W30)

PUBLICATIONS

Performance Tuning

- [arXiv 2024] Ke Cheng, Zhi Wang, Wen Hu, **Tiannuo Yang**, Jianguo Li, and Sheng Zhang, “Towards SLO-Optimized LLM Serving via Automatic Inference Engine Tuning”
- [ICDE 2024] **Tiannuo Yang**, Wen Hu, Wangqi Peng, Yusen Li, Jianguo Li, Xiaoguang Liu, and Gang Wang, “VDTuner: Automated Performance Tuning for Vector Data Management Systems”, in International Conference on Data Engineering (ICDE), 2024
- [ICPP 2023] **Tiannuo Yang**, Ruobing Chen, Yusen Li, Xiaoguang Liu, and Gang Wang, “CoTuner: A Hierarchical Learning Framework for Coordinately Optimizing Resource Partitioning and Parameter Tuning”, in International Conference on Parallel Processing (ICPP), 2023

Operations Research (Undergraduate Thesis)

- [JCR Q1] **Tiannuo Yang**, Zhongzhu Chu, and Bailin Wang, “Feasibility on the Integration of Passenger and Freight Transportation in Rural Areas: A Service Mode and an Optimization Model”, in Socio-Economic Planning Sciences, 2023.

Survey

- [arXiv 2023] .., **Tiannuo Yang** (Co-First Author), .. and Xiaodong Zeng, “On the Opportunities of Green Computing: A Survey” (Writing Section: 6.4 Resource Optimization)

INTERNSHIPS

Research Intern at UIUC, US

May 2024 ~ Present

Department of Computer Science, under Prof. Minjia Zhang

- System optimization for vector database and retrieval augmented generation (RAG)

- Research Intern** at Ant Group, Beijing, China June 2023 ~ Jan. 2024
Risk Intelligence Department, under Dr. Jianguo Li
- Optimizing large language model infrastructure - vector database
- Undergraduate Research Assistant** at UCAS, China Sep.2020 ~Sep. 2021
School of Economics and Management, under Prof. Guanghui Zhou
- Data-driven optimization of combinatorial optimization problem (vehicle routing) 2020

RESEARCH PROJECTS

- Qiyuan Laboratory Innovation Fund** Nov. 2023 ~ Present
College of Computer Science, Nankai University, under Prof. Yusen Li Core Member
- Resource isolation mechanism for multi-tenant cache system (Cachelib by Facebook)
- CCF-Ant Special Research Fund on Green Computing** Jan. 2023 ~ Jan. 2024
Ant Group, under Dr. Jianguo Li Core Member
- Online learning-based parameter tuning for OS Containers
- National Natural Science Foundation of China** Jan. 2023 ~ Present
College of Computer Science, Nankai University, under Prof. Yusen Li Member
- Improving resource utilization in the cloud with QoS guarantee
- Major Project of National Natural Science Foundation of China** Dec. 2022 ~ Present
College of Artificial Intelligence, Nankai University, under Prof. Xuebo Zhang Member
- Real-time scheduling of cluster robots for major equipment manufacturing

ACADEMIC COMPETITIONS

- 1st National Massive Storage Competition**, Huawei Technologies Co., Ltd. Dec. 2022
National Third Prize (RMB 10,000)
- Scheduling of Storage Data Retrieval Tasks
- 7th National ‘Internet+’ Innovation and Entrepreneurship Competition** Dec. 2021
Provincial (Tianjin) Gold Prize and nominated for the national finals
- Intelligent Tuning of Cloud Services (Cooperated with Huawei)
- National Contemporary Undergraduate Mathematical Contest** Sep. 2020
National Second Prize (top 2%)
- Mixed Integer Linear Programming and Heuristic Algorithms

TALKS

- CoTuner: A Hierarchical Learning Framework for Coordinately Optimizing Resource Partitioning and Parameter Tuning**
- Conference talk at the 52nd International Conference on Parallel Processing (ICPP), online, August 9, 2023
- VDTuner: Automated Performance Tuning for Vector Data Management Systems**
- Conference talk at the 40th International Conference on Data Engineering (ICDE), Utrecht, the Netherlands, May 17, 2024

HONORS

- 1st-Class Gongneng Scholarship** (RMB 12,000, top 10%), Nankai University 2023, 2024
- Top Ten Singers**, University of Science and Technology Beijing 2020