MUGE ZHANG

m.zhang1@student.fdu.edu https://mugezhangfdu.github.io/

EDUCATION

MS in Applied Computer Science, Fairleigh Dickinson University, GPA: 3.92/4.00 December, 2024 (expected)

Coursework: *Graduate Research*, *Artificial Intelligence*, *Big Data Analytics*, *Computer Algorithms*, *Database Systems*, *Systems Programming*, *Software Engineering*, *Assembly Language*, *Graduate Capstone Project*, *Linear Algebra (UCSD Extension)*, *Discrete Mathematics (UCSD Extension)*, *Calculus II (UCSD Extension)*, *Calculus I (ASU)*

BA in War Studies and Philosophy , King's College London, Upper Second-Class Honours June, 2021

International Exchange Program , Johns Hopkins University, GPA: 4.00/4.00 June, 2020

WORK EXPERIENCE

Binance US, **Backend Engineer Intern**, remote Sep. 2024 - present

- Developed and maintained internal backend tools using Python in a high-security environment

Wefind AI, **Software Engineer**, Beijing, China May 2021 - Aug. 2023

- Developed a machine translation workflow using Python, OpenAI API, and NLP libraries (spaCy, NLTK), improving translation accuracy by 30% and reducing manual post-editing by 50%
- Architected and deployed microservices with Flask, Docker, and Kubernetes on AWS, achieving 99.9% uptime and scaling the system to handle over 10,000 translation requests per day
- Enhanced system performance by implementing Redis caching and optimizing translation algorithms, reducing latency by 40% and increasing throughput by 25%

PUBLICATIONS

1. **Zhang, M.**, Lee, D. Y., Janarthanan, V., & Ryoo, J. (2024). **Microarchitectural Analysis of Pre-processing Stage in Machine Learning Workloads**. *7th International Conference on Algorithms, Computing and Artificial Intelligence*. IEEE. (Accepted)
2. Dai, W., & **Zhang, M**. (2024). **A High Performance AI-powered Cache Mechanism for IoT Devices**. *International Conference on Cyber-enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE. (Accepted)
3. Dai, W., & **Zhang, M**. (2024). **Hierarchical Agglomerative Clustering Optimization for Massive Data.** *IEEE 24th International Conference on Scalable Computing and Communications (ScalCom)*. (Accepted)

RESEARCH EXPERIENCE

Characterization of Neuro-Symbolic AI and Graph Convolutional Network workloads Oct., 2024 - present

Collaboration with Dr. Eugene B. John, Department of Electrical and Computer Engineering, UTSA

- Conducted in-depth analysis of Neuro-Symbolic AI (NSAI) and Graph Neural Networks (GNN) models, comparing performance characteristics with traditional deep learning models
- Identified limitations in parallelism for NSAI models due to complex control flow, low compute-to-byte ratio, and high data movement costs
- Analyzed graph networks, highlighting the prevalence of sparse matrix multiplication and similar low compute-to-byte operations, which reduce parallelism potential
- Proposed optimization techniques focused on improving element-wise operations for better computational efficiency in graph-based models

Energy Improvement Opportunities in Extended Reality (XR) Platforms Oct., 2023 - present

Gildart Haase School of Computer Sciences and Engineering, FDU Vancouver, Canada

Supervisor: Dr. Jeeho Ryoo

- Conducted a comprehensive micro-architectural analysis of different stages in XR and emerging ML workloads, identifying key bottlenecks in data-heavy applications across vision, audio, text, and multimodal domains
- Designed and implemented a bottom-up profiling methodology to uncover code hotspots, leading to insights that optimize the pipelines, reducing computational overhead and improving performance
- One paper accepted; Journal article in progress

Enhancing VLLM Performance through Cache Optimization May, 2024 - present
Gildart Haase School of Computer Sciences and Engineering, FDU Vancouver, Canada
Supervisor: Dr. Jeeho Ryoo

- Profiled and optimized the cache mechanism of VLLM, identifying performance limitations in the existing LRU algorithm used for caching
- Designed and implemented a new algorithm to replace the LRU approach, significantly improving VLLM's performance over the benchmark

Improving Hierarchical Clustering in Large-Scale Data Systems April, 2024 - present
Gildart Haase School of Computer Sciences and Engineering, FDU remote in Teaneck, NJ
Supervisor: Dr. Wenyun Dai

- Designed and implemented an optimization method for hierarchical agglomerative clustering within the MapReduce framework, reducing computational complexity by filtering marginal observations based on their centroid distances.
- Developed and tested corresponding MapReduce functions on multiple datasets, achieving a significant reduction in calculation volume and shuffle data, leading to 9% to 82.3% running time improvement while maintaining over 90% accuracy.

Cloud Integration and AI-Driven Caching for IoT Performance and Privacy April, 2024 - Aug., 2024
Gildart Haase School of Computer Sciences and Engineering, FDU remote in Teaneck, NJ
Supervisor: Dr. Wenyun Dai

- Designed and implemented a novel IoT framework where local devices act as the cache and remote Cloud servers function as memory, optimizing both performance and privacy
- Incorporated lightweight AI methods to predict user behavior patterns, enhancing system adaptability and maintaining optimal performance across diverse IoT applications

TEACHING EXPERIENCE

Veritas China Xi'an, China
Libral Arts Instructor June, 2019 - Aug., 2019

- Led group discussions for a class of 10-15 students, with instructor acceptance rate of approximately 5%

King's College London London, UK
Teaching Assistant Sep., 2020 - June., 2021

- Responsible for grading quizzes and commenting on papers for two first-year classes

HONORS

2024 **Graduate Academic Distinction Award**, Fairleigh Dickinson University Vancouver, Canada
2021 **Department Commendation Letter** , King's College London London, UK

SKILLS

- C/C++, Python, Matlab, Latex
- Pytorch, Keras, Cuda
- Windows, Linux