ZUNYAO MAO

maozy2018@mail.sustech.edu.cn · (+86) 15813849498

Education

Southern University of Science and Technology, Shenzhen 2018 – 2022

- *Bachelor Degree* Computer Science and Technology
- *GPA: 3.82* 15/160, top10%

Southern University of Science and Technology, Shenzhen 2022 – Present

- *Mphil.* Computer Science and Technology

Publication

gSampler: General and Efficient GPU-based Graph Sampling for Graph Learning.

Ping Gong, Renjie Liu, **Zunyao Mao**, Zhenkun Cai, Xiao Yan, Cheng Li, Minjie Wang, Zhuozhao Li.
Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), Koblenz, Germany, October 2023.

Speeding Up End-to-end Query Execution via Learning-based Progressive Cardinality Estimation.

Fang Wang, Xiao Yan, Man Lung Yiu, Shuai Li, **Zunyao Mao**, Bo Tang.
Proceedings of the ACM Conference on Management of Data (SIGMOD), Seattle, WA, USA, June 2023.

GHive: Accelerating Analytical Query Processing in Apache Hive via CPU-GPU Heterogeneous Computing

Haotian Liu, Bo Tang, Jiashu Zhang, Yangshen Deng, Xinying Zheng, Qiaomu Shen, Xiao Yan, Dan Zeng, **Zunyao Mao**, Chaozu Zhang, Zhengxin You, Zhihao Wang, Jiang Runzhe, Fang Wang, Man Lung Yiu, Huan Li, Mingji Han, Qian Li, Zhenghai Luo.
Proceedings of the ACM Symposium on Cloud Computing (SoCC), San Francisco, California, November 2022.

Award

First Prize,ASC Student Supercomputer Challenge 2022 March

Team Member

Large Scale Distributed Training Problem : Deep-MD, a multilayer perceptron model for molecular dynamics simulation based on TensorFlow. Yuan1.0, a large-scale Chinese language pretraining model.

- Using Nsight System, a system analysis was performed to identify the performance bottlenecks of Deep-MD running on TensorFlow.
- By utilizing methods such as **XLA** and **computational graph optimization**, achieved a 50% acceleration effect on training data for DeepMD.
- Utilized various distributed training communication libraries like Horovod and DeepSpeed.

Southern University of Science and Technology Outstanding Student Scholarship. 2019, 2020, 2021

- Continuously ranked in the top 10% for overall performance for four consecutive years.

Internship and Project Experience

Amazon Web Services AI Lab, DGL Group July 2022 – Present

Research Internship Python, C/C++, CUDA

Developed and improved **graph sampling** module for DGL, a widely-used GNN training system.

- Surveyed 10+ classical or state-of-the-art graph sampling algorithms.

- Used torch.fx to abstract the intermediate representation (IR) of graph sampling and implemented low-level kernel fusion optimization in CUDA.
- Performed baseline experiments and comparisons across multiple systems, including PyG, DGL, etc.
- The related papaer was accepted by SOSP 2023.

Huawei 2012 Lab, Central Software Institute, Gauss Database February 2022 – June 2022

Industrial Internship C/C++

Developed the underlying **distributed communication** module for a cloud-native database, utilizing **RDMA network programming**.

- Participated in experiments and developed different IO models for multithreaded communication in high-concurrency scenarios.
- Designed and implemented a **dual-buffer data structure** for read-write separation in high-concurrency environments.
- Contributed to the design and development of the underlying RDMA communication.

GHive June 2020 – November 2021

Resear Project Java, C++, CUDA

A distributed CPU-GPU heterogeneous computing data processing engine based on Apache Hive, accepted by SoCC 2022.

- Participated in the implementation of the Map-side GPU interface for the Tez execution engine.
- Implemented SQL operators like Hashjoin using **CUDA**, achieving up to 100x acceleration at the operator level.
- Achieved 2-3x end-to-end acceleration in the SSB international benchmark test on an 8-node server cluster.

LPCE March 2021 – June 2021

Reseach Project Python, PyTorch

A machine learning-based algorithm for SQL query cardinality estimation, achieving end-to-end acceleration in SQL execution on the open-source PostgreSQL database.

- Participated in the development and implementation of LSTM models using **PyTorch**.
- Participated in model training experiments and optimization. LPCE achieved an average acceleration of 86.8
- Contributed to the porting process of the new cardinality estimation algorithm to the Postgres database, enabling re-optimization of SQL execution plans.

Texera June 2021 – September 2021

UC Irvine Summer Research Project Scala, Java

A visual streaming SQL engine.

- Designed and implemented the Interval Join operator.
- Designed and implemented distributed performance monitoring based on an Actor system.

DTOP April 2021 – June 2021

Operating Systems Course Project C++

A distributed memory monitoring system that detects memory usage and potential memory leaks in a cluster of processes.

- Provided a web-based UI interface to monitor the memory usage of each node in the cluster.
- Utilized gRPC for inter-node communication.
- Implemented memory usage statistics by dynamically replacing the malloc function.

IT Skills

- Programming Languages: Java, C/C++, Python, JavaScript, SQL, etc.
- Frameworks: MPI, Tensorflow, PyTorch, CUDA, MPI, SpringBoot