Guoyu Chen

Columbus, OH
(928)310-5569
chen.9605@osu.edu

Summary

Researcher specializing in system design and performance optimization for data center GPU servers. Proven track record of designing effective control frameworks to enhance the processing efficiency of Machine Learning/GenAI workloads on servers. Equipped with strong problem-solving skills to optimize the system performance of Machine Learning/GenAI infrastructure based on workload analysis.

Research Experience

Graduate Researcher/Research Associate 01/2021 – 05/2024

The Ohio State University, Columbus, OH

- Research on power-aware computing systems including data center servers and GPUs
- Proposed control frameworks to improve the processing efficiency of ML workloads

Co-location of ML Workloads on Data Center Servers:

- Exploited Multi-Process Service to enable GPU spatial sharing of training and inference workloads
- Designed a two-tier control framework to allocate GPU resources dynamically
- Explored contextual multi-armed bandits as the GPU resource optimizer
- Ensured inference workloads with stringent time requirements to meet Service-Level Objectives
- Built a small-scale data center GPU systems
- Saved data center capital expense up to 74.9%
- Accepted in ICDCS 2024

Power and Latency Control of Data Center Servers Running ML Workloads:

- Designed a LQR MIMO controller to adjust power and ML inference latency; The proposed control framework caps server power and ensures meeting latency constraints of ML applications
- Enabled safe oversubscription of power facilities in data centers by power capping to avoid overload and overheating
- Trained ML models by manipulating model width with a single knob slice rate
- Implemented Resnet models with model slicing to process images using PyTorch, TensorFlow, and MXNET
- Can be applied in Generative AI systems
- Presented and published in ICDCS 2022; Accepted by journal ACM TAAS in 2024

Professional Experience

Software Engineering Internship 06/2021 - 08/2021

Schweitzer Engineering Laboratories, Inc., Lewis Center, OH

- Processed three-phase voltage/current signal of power transmission lines using FFT and implemented with C++ and CUDA
- Designed a system displaying time-series data with InfluxDB
- All implementations are through fiber-optic and Ethernet communication between Nvidia Jetson AGX Xavier and the power line protection device

Machine Learning Course Project 11/2020-12/2020

- Classified a given tabular dataset with 20 classes, 20 features and 100, 000 samples
- Designed a SVM classifier and a Deep Neural Network model using Pytorch
- Collaborated with a team of three to design Random Forest and XGBoost classifiers
- Worked with a team of three to write a report, leading to the highest inference accuracy in class

Skills

Data Center Servers, Control theory, Machine learning, Python, System design, GenAI, Computer networks

Education

Electrical and Computer Engineering

The Ohio State University (GPA: 3.87/4.00) 05/2024

Master of Science (2019-2020) and Doctor of Philosophy (2021-2024)

Electrical Engineering

Northern Arizona University (GPA: 4.00/4.00) 05/2019

Bachelor of Science

Chongqing University of Posts and Telecommunications (GPA: 3.85/4.00)

Bachelor of Engineering 07/2018

Teaching Experience

Graduate Teaching Associate 08/2021-12/2023

- Ohio State ECE2060 Digital Logic course coordinator in Au21
- Ohio State ECE3567 Microcontroller lab instructor in Sp22, Au22, Sp23 and Au23

Selected Publication

- **Guoyu Chen** and Xiaorui Wang. 2024. OptimML: Joint Control of Inference Latency and Server Power Consumption for ML Performance Optimization. ACM Trans. Auton. Adapt. Syst. Just Accepted (May 2024). https://doi.org/10.1145/3661825
- **Guoyu Chen**, Srinivasan Subramaniyan and X. Wang, "Latency-guaranteed Co-location of Inference and Training for Reducing Data Center Expenses," 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS), Jersey City, New Jersey, USA, 2024 (Accepted to appear).
- **Guoyu Chen** and Xiaorui Wang, "Performance Optimization of Machine Learning Inference under Latency and Server Power Constraints," 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS), Bologna, Italy, 2022.
- Yunhao Bai, **Guoyu Chen**, and Xiaorui Wang, "Fusing WiFi Signals and Camera for Driver Activity Recognition based on Deep Learning", the 19th IEEE International Conference on Mobile Ad Hoc and Smart Systems (MASS 2022), Denver, Colorado, October 2022.