MINJIE MAO

Student Pursuing Master's Degree

@ minjiema@usc.edu   ☎ +1 (734) 604-5631   ✉ 1420 W 22ND ST, Los Angeles, CA 90007   defthobo   in LinkedIn

EDUCATION

University of Southern California — Los Angeles, USA
Master of Science in Computer Science | GPA - 3.67/4.00 — Sep 2023 – May 2025 (Expected)

University of Michigan - Ann Arbor — Ann Arbor, USA
Bachelor of Science in Computer Science (Dual Degree) | GPA - 3.78/4.00 — Sep 2021 – Apr 2023

Shanghai Jiao Tong University — Shanghai, China
Bachelor of Science in Electrical and Computer Engineering (Dual Degree) | GPA - 3.62/4.00 — Sep 2019 – Aug 2023

TECHNICAL SKILLS

Languages: Python, C/C++, SystemVerilog, JavaScript, TypeScript, Java, Kotlin, Bash Scripting, HTML, CSS
Frameworks & Development Tools: Linux, CUDA, NodeJS, React, Angular, Bootstrap, NeuronxDistributed
Databases: SQLite, Oracle, MongoDB

WORK EXPERIENCE

Amazon Web Services — San Jose, USA
Software Dev Engineer Intern - Machine Learning Chip Accelerator | NeuronxDistributed, LLM — May 2024 – Aug 2024

- Implemented multi-modal model inference on AWS ML chips (Trn1/Inf2) under NeuronxDistributed framework.
- Enabled input and output channel level tensor parallel convolutional layer on neuron device.
- Designed preprocessing of the embedding mask on CPU to avoid tensor materialization and IR graph truncation.

Shanghai Jiao Tong University — Shanghai, China
Engineering Drawing OCR Systems Research Assistant | Python, PaddleOCR, PyTorch — Dec 2022 – Nov 2023

- Augmented data according to test results, including the generation of annotations with complex background.
- Added triplet vector embedding techniques to PaddleOCR workflow, improving recognition accuracy to over 80%.

PROJECTS

Responsive Web & Android App Dev | Typescript, NodeJS, Angular, MongoDB, Kotlin, REST API — Jan 2024 – May 2024

- Developed a real-time stock search website with responsive design using Bootstrap and Angular framework. ↗
- Deployed NodeJS backend server and MongoDB Atlas to persistently store user information.
- Enabled gesture control features on migrated android stock search application.

Reliable Transport Protocol & BBR Congestion Control | C/C++, Socket Programming, Quagga, BBR — Sep 2023 – Dec 2023

- Built a reliable transport protocol on top of UDP with cumulative acknowledgment and sliding window to resolve issues of packet loss, delay, corruption, duplication, and reordering.
- Configured OSPF, iBGP and eBGP within an Autonomous System using Quagga routing suite.
- Added Bottleneck Bandwidth and Round-trip propagation time(BBR) congestion control protocol to simplified TCP.

R10K-style Out-of-Order Superscalar Processor | SystemVerilog — Jan 2023 - Apr 2023

- Oversaw the intricate design of critical components of an R10k-style out-of-order superscalar processor, including the Reorder Buffer (ROB), Reservation Stations (RS), and Physical Register File (PRF) to increase instruction level parallelism.
- Implemented a multiport non-blocking instruction cache and an instruction fetcher coordinating with branch predictor.
- Engineered a 2-way set-associative data cache featuring a victim cache under LRU policy, to minimize cache misses.

Fakebook Database | Oracle SQL, MongoDB, Java, JavaScript — Jan 2023 – Apr 2023

- Designed an ER Diagram and built a schema of tables from a Facebook-like application database.
- Constructed SQL queries using Java and JDBC to extract useful information from fakebook database.
- Migrated tables in fakebook database to a MongoDB collection of users and finished queries on the collection.

Optimization of the Neural Network Convolutional Forward Pass | CUDA C/C++ — Sep 2022 – Dec 2022

- Adopted shared memory tiling technique to reduce global memory traffic of the forward path of convolutional layers.
- Reduced convolution kernel to a highly efficient matrix multiplication kernel with input unfolding and replicating.
- Exploited multiple fine-tuned kernels tailored for various convolutional layers within deep learning models.

Operating System Emulation | C/C++, Multithreading, Network File System — Sep 2022 – Dec 2022

- Customized a thread library with mutex support to manage uniprocessor scheduling.
- Implemented virtual memory management pager, optimizing memory allocation and resource utilization.
- Orchestrated emulation of a multi-threaded, multi-client, and secure network file system server, minimizing disk I/Os.