




Daniyal Khan

 danikhan632 |  daniyalmkhan/ |  danikhan632@gmail.com | US Citizen | www.daniyalkhan.dev/

EDUCATION

Georgia Institute of Technology

B.S. Computer Science

December 2023

High Honours

Concentrations: Intelligence/AI and Systems and Architecture

Relevant Coursework: Agile Development, Artificial Intelligence, Advanced Algorithms and Data Structures, Robotics and Perception, Computer Architecture, Circuit Design Lab

WORK EXPERIENCE

Samsung Semiconductor

December 2024 - Present

Senior Engineer, AI/ML Software Compiler

San Jose, CA

- Designed and implemented a custom backend for the Triton compiler, enabling efficient deep learning model execution on specialized hardware platforms by leveraging advanced optimization techniques and integration with LLVM-based infrastructure.
- Developed and optimized a custom LLVM target to align with unique hardware architectures, introducing innovative solutions for staged lowering and data locality

Manhattan Associates

January 2024 - December 2024

Software Engineer

Atlanta, GA

- Developed microservices for shipment and transportation order management using Java and Spring Boot, enabling real-time tracking and efficient allocation of trailers, equipment, and resources
- Implemented a Transportation Resource Group (TRG) algorithm to optimize shipment routing by matching transportation modes (ocean, air, rail), carriers, and equipment, enhancing logistics efficiency.
- Enhanced shipment rate calculation performance by implementing a prefetching mechanism to load data on carriers, facilities, and warehouses, etc in advance.

NCR Software Engineering

May 2022 - August 2022

Software Engineering Intern

Atlanta, Georgia

- Led the creation of an internal debugging tool, facilitating real-time monitoring and management of MQTT messages
- Designed and implemented a dynamic frontend using React, deeply integrating TypeScript and Redux to ensure a seamless user experience and efficient state management.
- Mastered the intricacies of SQL to ensure optimal logging, storage, and retrieval of MQTT messages, enhancing system responsiveness and reliability.
- Pioneered a custom TreeSet data structure, optimizing data modification and retrieval processes

EXPERIENCE & OPEN SOURCE CONTRIBUTIONS

Triton Compiler & Runtime

Dec 2023-

- Designed Triton backend for VortexGPU, including MLIR passes for shared memory allocation, async bulk memory operations, dot operations and a pytorch memory allocator for global memory tensors
- Enhanced Visualization tool featuring step-by-step kernel debugging, animated tensor operations, and heat map representations, while integrating line trace data and performance metrics to provide comprehensive analysis of 2D and 3D tensor operations within a debugger-style interface.
- Developed MLIR passes in C++, optimizing GEMM through tiling and vectorization, and decomposing them into outer products for enhanced performance on Arm Scalable Matrix Extension processors.

LlamaGym

May 2024 -

- Developed a reinforcement learning environment to enhance LLM reasoning capabilities by training agents to solve Math and Logic problems, utilizing scalar rewards to train LLMs
- Implemented CoT training setup involving Critic agents on semi-supervised text datasets
- Designed reasoning token placement based of entropy and varentropy of predicted tokens for enhanced LLM reasoning

PROJECTS

- GPU Programming & Architecture Lectures** Delivered a lecture on GPU programming and architecture, covering kernels, threads, blocks, grids, shared memory, and tensor cores, to an audience of computer science students.
- OpenAI Finetuning Server** - Developed a scalable, OpenAI-compatible server for finetuning LLMs, integrating Redis and an asynchronous architecture to support distributed training across multiple nodes.
- nanoLlama** - Developed a streamlined repository for efficient training and fine-tuning of medium-sized LLaMA-2 models, similar to nanoGPT for GPT-2
- AutoGPT plugins** - Developed AutoGPT plugins that leveraged large language models (LLMs) to enable automated stock trading via the Alpaca Trading API and facilitate AI-powered text message communication

SKILLS

Programming languages: Python, C++, Java, C, JavaScript, Rust

Frameworks Software: MLIR , Docker, LLVM, CUDA, Pytorch, React, Springboot, Maven