Daniyal Khan

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

Education

Georgia Institute of Technology December 2023
B.S. Computer Science High Honors

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

Experience & Open Source Contributions

Triton Compiler & Runtime July 2023-

- Developed MLIR (Multi-Level Intermediate Representation) passes in C++, optimizing GEMM through tiling and vectorization, and decomposing them into outer products for enhanced performance on Arm Scalable Matrix Extension and Scalable Vector processors.
- Engineered prefetching strategies within the compiler to improve the efficiency of arithmetic units, addressing and mitigating the limitations of CPU memory bandwidth.
- Resolved build pipeline challenges for open-source repositories including Triton and Triton-shared, through the integration of cross-compiling LLVM binaries for arm64 and integrated test runners for arm64
- Designed MLIR lowering strategies for tensor reduction, and matrix multiplication to improve the performace of Large Language Model infrence

Guidance LLM inference server May 2023

- Enabled support for diverse Large Language Models (LLMs) inference engines, facilitating Quantized LLMs to generate structured output, enhancing the accuracy and usability of model predictions.
- Developed an OpenAI-compatible inference server leveraging Logit and Regex processors, featuring structured JSON output and function calling capabilities, to streamline integration and expand functionality.
- Innovated the token generation workflow by integrating operations with a REST API, optimizing performance and scalability by offloading processing.

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

Georgia Tech Vertically Integrated Project January 2022 - May 2023
Team Phoenix-High Performance Computing Atlanta, Georgia

- Developed high-performance computing (HPC) applications by writing and optimizing CUDA kernels, leveraging NVIDIA GPU architectures to achieve significant improvements in computational efficiency and throughput.
- Applied advanced parallel computing techniques and memory management optimizations within CUDA, significantly improving the performance of LINPACK's dense linear algebra computations.
- Engineered a bf16 Tensor Core GEMM (General Matrix Multiply) kernel, leveraging NVIDIA's Tensor Cores to accelerate mixed-precision computations, for Machine Learning and AI workloads

Projects

LLama-gym: Reinforcement Learning Environment Trainer 2024

- Implemented Reinforcement Learning environments for LLM agent to solve Math, Coding , and Logic based problems rewarding agents for actions that moved closer to goal state
- Architected the system using the Ai agent framework to enchance LLM reasoning capabilites
- Created mechanism to utilize output dato to finetune LLMs using Proxminal Policy Optimization and other RLHF techniques

BuzzOS 2023

- Implemented and optimized entire userspace and userspace libraries, streamlining system calls and facilitating seamless interaction between userspace processes and kernel functionalities, thereby enhancing overall system efficiency and usability.
- Designed and integrated user libraries tailored to the system's specific requirements, providing an intuitive interface for invoking system calls from userspace processes, thereby simplifying application development and enhancing system accessibility.

Skills

Programming languages: Rust, C++, Go, Python, Java, C#, SQL, Bash, JavaScript, HTML, CSS
Frameworks Software: MLIR , Docker, LLVM, CUDA, Pytorch, React, Springboot, Maven, React Native