GOKULKRISHNA MUTHUSAMY

+1 201-241-5854 | New York City | gm3314@nyu.edu

Graduating May, 2025. Looking for Full time opportunities [LinkedIn] [Website][Github]

Education

New York University (Courant), New York, NY — Sept 2023 - May 2025
Master of Science in Computer Science — 4.0/4.0
National Institute of Technology (NIT), Tiruchirappalli, TN — Jun 2020
Bachelor of Technology in Computer Science and Engineering — 8.6/10

Experience

Machine Learning Intern, **AMD,** San Jose, CA — May 2024 – Aug 2024
Masters Tech – AI Engine Driver Team, AMD Adaptive Edge Computing Team.

- Developed prototype compiler using MLIR and Clang Frontend, simplifying AI Engine (NPU) programming by emitting optimized driver APIs for Versal Adaptive SoCs. The tool captures AI Engine driver parameters and data structures using ClangAST and uses MLIR (Types, Attributes and Operation) to emit code.
- Fixed the Routing API for NPU Tiles by resolving errors, and bugs and writing unit tests for edge cases. This resulted in the Upstreaming of the API for the AI engine runtime driver codebase for Versal Adaptive SoC.
- Researched Hybrid programming and Multi-language support using compilers and IR.

Senior Software Engineer, Samsung Research India, Bangalore, KA — Apr 2022 – Aug 2023
SNAP - CPU, GPU and Compiler team for Neural Acceleration - OnDevice AI.

- **Accelerating Generative AI models**: Programmed OpenCL GPU kernels for complex neural layers like group normalization and updated MLIR to support TfLite model conversion from Keras, enabling complete execution of Stable Diffusion models on Mobile GPUs. Engineered optimizations like convolution splitting, Quantization, GELU approximation, and fused softmax in VAE, which resulted in a 4.58x improvement in performance over CPU. This solution will serve as a base for infrastructure development for the acceleration of Gen-AI models on Samsung Mobile GPUs.
- **Automatic Caching tool**: Designed an automatic recompilation & caching tool for OpenCL kernels on GPUs to reduce caching overhead and prevent crashes during GPU driver updates, which resulted in a 25% reduction in crash reports and a 20% increase in the productivity of engineers and the use case team. Received Spot award in Q3 2022 for developing this tool.
- Accelerated 10 use cases on Samsung flagships (Galaxy S23) to improve user experience, achieved a 20% improvement in load time and 10% in execution time and overall performance improvement of 1.4x (S23) and 1.3x (Fold5) using ML Accelerators.

ML Software Engineer, Samsung Research India, Bangalore, KA — Jan-2021 – Apr 2022
SNAP - CPU, GPU and Compiler team for Neural Acceleration - OnDevice AI.

- **Framework inference Profiler**: Engineered a Profiling tool for calculating the layer-wise performance of kernels in ArmNN during inference of ML model. Resulting in the faster diagnosis of performance degradation and improved the productivity of the team by 30%.
- Collaborated with Galaxy RAW (USP) team to solve greenish tinge issues on images from AINR (noise reduction) by implementing new normalization method to support float16 quantized precision execution. This improved inference time by 34% over the previous method.
- Directly helped in enabling over 15 USP camera and gallery features deployed on Samsung Galaxy S22. Contributing to over 1.5x optimizations for speed, memory and battery.

Intern, ***Samsung Research India***, Bangalore, KA — May 2019 – July 2019
Keyboard Intelligence - OnDevice AI

- Improved the emoji prediction in the Samsung keyboard by incorporating keystroke statistics in the NLP sentiment analysis algorithm, resulting in a 5% improvement in accuracy on personalized datasets.

Technical Skills

Programming Languages: C++, Python, OpenCL, CUDA, C, JavaScript, Shell, Rust, HTML, CSS
Technologies: TensorFlow, Pytorch, TFLite, Keras, ArmNN, ONNX, MLIR, Apache TVM, MongoDB, RapidMiner, PostgreSQL, Quantization, LLVM, NPU programming, Clang AST & Frontend.
Tools: Git, Perfetto, Gdb, Vitis, FlatBuffers, Bazel, Valgrind, Asan, Hwasan, Android NDK, Perforce,

Projects

Monkey Language Interpreter in C++ (project link) Dec, 2024

- Implemented monkey language interpreter from scratch inspired by Thorsten Ball's book "Writing Interpreter in Go". This involved generating AST using a recursive descent parser (using Pratt parsing), an object system & environment to bind variables and REPL to execute code.
- The language supports integers, infix operators, prefix operators, variables, functions, parsing errors and conditional statement.

GGlow – Experimental Graph Lowering IR (WIP) (project link) Aug, 2024

- Graph lowering (Glow) experimental implementation using MLIR. The project attempts to load the model using torchscript c++ API, parse it into GGlow IR (Graph Lowering IR using MLIR) then lower it to llvm-dialect or llvm-ir and execute it using OrcJIT execution engine. The aim is to create simple API to lower any pytorch model using torchscript.

Trajectory Learner – Model Predictive control (MPC) using JEPA (project link) Dec, 2024

- Implemented and trained a JEPA architecture MPC on a dataset of 2.5M frames of exploratory trajectories. The model generated representations that captured the true coordinates of the agent. The representations were not created using generative models, but using the JEPA model using VicReg loss function and accuracy tested via linear probing.

Patents & Publications

A1 graded patent by Samsung Research HQ Oct 11, 2022

OnDevice Validation: The main inventor of Systems and Methods for On-Device Validation of a Neural Network Model, this idea reduces validation memory requirement by 100% and makes the computations on-device friendly. Inventors- M Gokulkrishna, Siva Kailash, R. Prasanna, Rajath Elias, Ashok Kumar, Praveen. **Pending patent**: US-20240135181-A1