Kushhal Sathish

+1-408-591-1725 | kushhalsathish1911@gmail.com | Linkedin: kushhalsathish | Projects

EDUCATION

Santa Clara University **Jan 2024 - Dec 2025**
Master of Science, Computer Science and Engineering

Anna University, MIT Campus **Aug 2018 - Apr 2022**
Bachelor of Technology, Information Technology

SKILLS SUMMARY

- **Languages**: C, C++, Python, Java, RISC-V
- **Compiler & Frameworks**: LLVM, MLIR, TensorRT, vLLM, Triton
- **Kernel Development**: CUDA, ROCm, Triton, Pallas, FlashInfer, JAX
- **ML frameworks**: PyTorch, TensorFlow
- **Technical Skills**: Deep Learning, Generative AI, Compiler Optimizations, Computer Architecture, Database systems , Parallel/concurrent Programming, High Performance Computing, Operating Systems, Distributed Systems, assembly programming, Data Structures and Algorithms, version control, Probabilistic programming, Hardware-Software Codesign

EXPERIENCE

Santa Clara University - Research Developer **Feb 2025 - Nov 2025**
- Designed an adaptive GPU cache bypassing technique using Program counter predictors in **C++** to enhance L1-D cache efficiency. Leveraged **Accel-Sim** for simulation, achieving power and performance improvements by reducing cache pollution.
- Validated and benchmarked, optimized cache hit ratios, and resolved complex integration issues using **gdb**.
- Achieved **13% performance** improvement over a standard 16KB L1 cache.Reduced L1 cache **energy consumption by 25%** on average.Cache misses are reduced by **58%** on average.

Research Developer **Aug 2024 - Dec 2024**
- Developed a static analysis tool for LLVM IR which computes floating point rounding error bounds for basic blocks and implemented pipeline to estimate error propagation across input intervals with over 10000 operators.
- Implemented analysis feature for accuracy loss during precision downcasting.

Anna University **Mar 2023 - Nov 2023**
Research assistant *TamilNadu, India*
- Developed an Human Object interaction detection model using the transformer based encoder-decoder, trained on **V-COCO and HICO-DET** image dataset and enhanced the accuracy and efficiency of detecting and interpreting interactions between humans and objects.
- Achieved **inference time under 1 ms**, being significantly faster than parallel detectors (5-9 ms).

TEKION Corp **Jan 2022 - Jan 2023**
Associate Software Engineer - Intern and FTE *TamilNadu, India*
- Collaborated on advanced features for a chatbot, leveraging research to enhance user experience and quantitative evaluation of NLP models.
- Developed the narrative framework for chatbot, incorporating user interactions and dialogue paths to optimize data labeling and feature extraction.
- Enhanced intent and entity recognition, resulting in a **97%** increase in accuracy and model fine-tuning in the chatbot system.
- Created modules such as NLU Feedback module, Verbatim Bot Response Editor, NLU monitoring inbox, NLU Data Generator module,NLU Pipeline and Data Segregation.

PROJECTS

JIT Optimization for GPU Kernels
- Designed a JIT compilation framework with caching mechanisms for GPU kernels operating directly on LLVM . Implemented a lightweight annotation system to mark GPU kernel functions, allowing to extract IR and runtime variables during AOT compilation.
- Developed a runtime specialization engine that propagates argument values and thread launch parameters into GPU kernels, dynamically recompiling optimized variants on the fly. Increased performance up to **2.8× speedup** on AMD GPUs and **1.78×** on NVIDIA GPUs

Scalable Weakly-consistent Infection-style Process Group Membership Protocol
- The project was developed using Java to create a communication mechanisms in large-scale multi-region distributed systems in the cloud network using gossip protocol.
- Implemented failure detection using indirect ping mechanism, Bully Leader Algorithm, Cluster formation and communication, Gossip dissemination, Membership Protocol, Data recovery.