

Sai Saketh Cherukuri

GPU Computing · ML Systems · High Performance Inference | 14502 Dowitcher Way, Boyds, MD-20841

saisakethch.com | github.com/ssaketh-ch | LinkedIn/sai-saketh-cherukuri

Authorized to work in the United States without sponsorship (U.S. Permanent Resident).

Contact: +1 571-969-9304 | ssaketh.2002@gmail.com

ABOUT

Machine Learning Systems Engineer focused on LLM inference, performance analysis, and high throughput serving systems. Hands on experience with MLPerf Inference benchmarking, analyzing throughput bottlenecks, KV cache behavior using a **self developed** CLI based KV cache analysis tool, and scheduling trade offs in large language model workloads. Familiar with GPU architecture, profiling, and performance analysis using Nsight Systems and Nsight Compute, with practical experience studying transformer inference and serving frameworks such as vLLM. Interested in building scalable, cost efficient AI infrastructure and improving serving performance through measurement driven optimization.

EDUCATION

M.Sc. Mathematics (Computer Science Specialization) · Sri Sathya Sai Institute of Higher Learning 2024 - 2026

- Grade : 8.0/10.0 (O - Outstanding)
- Coursework: , Design and Analysis of Algorithms, Computer Networks, Systems Programming, Operating Systems, Computer Architecture, , Cryptography, Database Management, C++ Programming , Java Programming , Advanced Linear Algebra, Probability & Statistics, Optimization Techniques

B.Sc. (Hons.) Mathematics · Sri Sathya Sai Institute of Higher Learning 2020 - 2023

- Grade : 7.2/10.0 (A+)
- Coursework: Data Structures and Algorithms, Basic Computer Architecture, C Programming, Linear Algebra, Numerical Analysis

PROJECTS

MLPerf Llama 3.1 8B Inference Benchmarking · M.Sc Dissertation - NVIDIA H200 - LLM Inference

- Designed and built a fully reproducible benchmarking harness for the MLPerf inference suite targeting a single H200 GPU running Llama 3.1-8B
- Profiled kernel execution with Nsight Compute and Nsight Systems to trace compute vs. memory bandwidth bottlenecks at each transformer layer
- Swept quantization configs (FP8, BF16), KV cache budgets, and scheduling strategies, then documented the throughput and accuracy surface across conditions.
- Published code on GitHub alongside a detailed technical report covering methodology, raw numbers, and tuning recommendations
- Stack: Python, vLLM, Nvidia Nsight profilers, MLflow, Docker

H200 Quantization Suite · NVIDIA H200 - Quantization & Inference - Accuracy vs Performance

- Built a quantization benchmarking suite for evaluating FP8, INT8, and BF16 precision formats on an NVIDIA H200 GPU, targeting LLM inference workloads
- Measured throughput, latency, and accuracy trade offs across quantization schemes to identify optimal configs for cost efficient serving
- Documented findings in a reproducible format with raw numbers, methodology published on GitHub with the dataset on HuggingFace
- Stack: Python, CUDA, vLLM, Docker, Hugging Face

Periocular Age & Gender Recognition · B.Sc Project - Deep Learning - PyTorch

- Built an end-to-end periocular biometric pipeline for demographic prediction, from raw image preprocessing through evaluation on held out sets
- Compared several CNN backbone architectures on cropped eye-region data, balancing accuracy against inference speed and model size
- Wrote a structured technical report covering dataset curation, augmentation strategy, architecture rationale, and ablation results
- Stack: Python, PyTorch

Byte Inspector: Static Analysis Toolkit · M.Sc. Course Project - Java

- Developed a static analysis toolkit that flags common code quality issues across C++, CUDA, and Python source files
- Built a JavaFX desktop interface allowing developers to browse findings interactively, configure rule severity, and export annotated reports
- Reinforced understanding of parsing, AST traversal, and cross-language tooling and computer networking, relevant to compiler and developer tools work
- Stack: Java

SIC/XE Assembler · Academic Project - C++ - SIC/XE assembly

- Implemented a two-pass SIC/XE assembler in C++, handling symbol table construction, directive resolution, format-1 through format-4 instructions, and standard object record emission
- Strengthened understanding of the assembler/linker interface and how instruction encoding decisions affect binary size and relocation

Recursive Descent Parser (Pascal like Language) · Academic Project - C - SIC/XE assembly

- Wrote a recursive descent parser in C for a simplified Pascal like grammar, covering declarations, assignments, arithmetic expressions, and if/while control flow.
- Practical grounding in low level parsing, parse tree construction, and error recovery, directly applicable to compiler front ends and language tooling.

TECHNICAL SKILLS

Languages: C · C++ · Python · Java

ML / DL: PyTorch · MLPerf inference suite · MLflow · HuggingFace

GPU & HPC: CUDA · vLLM · Nsight Compute · Nsight Systems · kernel profiling & tuning · SGLang · TensorRt-LLM

Cloud & Infra: AWS related services · Kubernetes (Linux Foundation)

Systems & Tooling: Linux · Git · Docker · SLURM

Other: SAS Base Programming · JavaFX

CERTIFICATIONS

- AWS Certified Solutions Architect - Associate · Amazon Web Services
- AWS Certified Developer - Associate · Amazon Web Services
- AWS Certified Cloud Practitioner · Amazon Web Services
- SAS Certified Specialist: Base Programming · SAS Institute

OPEN SOURCE & GITHUB

All projects are publicly available and documented on GitHub at [ssaketh-ch](https://ssaketh-ch.github.io). Repositories include the MLPerf benchmarking harness, the periocular recognition model, the Byte Inspector toolkit, the SIC/XE assembler, and the recursive descent parser. Code is written with readability and reproducibility in mind.

BEYOND THE TERMINAL

I tend to approach hobbies the same way I approach systems work by trying to understand the fundamentals before reaching for shortcuts. I'm an avid photographer, you can find some clicks at : ssaketh-ch.github.io/photography-portfolio

- Sports - Cricket, Soccer, Badminton | Musician - Tabla player, Drummer - Photography, Videography and editing
- Elected Campus Sports Captain (2023): organized intra campus athletics, and inter campus sports activities.