Sakshi Sindhwal

M.Tech CSE
Indian Institute of Science, Bangalore

+91-9808109382
sakshi.dakshana16@gmail.com
ssakshi@iisc.ac.in
LinkedIn Profile

Education

- **Indian Institute of Science, Bangalore** *2023-25(current)*
  M.Tech , Computer Science Engineering CGPA: 8.0
- **National Institute of Technology, Uttarakhand** *2017-21*
  B.Tech , Electronics and Communication Engineering CGPA: 9.26
- **Jawahar Navodaya Vidyalaya, Dehradun** *2016*
  Central Board of Secondary Education Percentage: 93.2%

Personal/Coursework Projects

- **LLM Inference Optimization at High Performance Computing Lab.** *[2024-25]*
  M.Tech Project
  - The primary objective is to reduce the inference latency of large language models by minimizing the memory overhead associated with KV caches and reducing the KV cache copy overhead on both GPUs and CPUs.
  - Developing a hybrid framework that mitigates the drawbacks of Attention Sink (KV cache copy overhead) and Balancing Memory and Compute (memory overhead) for both CPUs and GPUs.
- **Building an in-kernel, per-process sandbox** *[2024]*
  LLVM , eBPF , Operating Systems , System Security
  - In this project we build a kernel module that accepts a per-process policy regarding what calls (library calls in this case) the process can execute (and at what points during its lifetime using it can execute those calls) using writing the LLVM passes, and enforce that policy within the kernel with the help of eBPF .
- **Microservice Implementation for Booking System using Spring** *[2024]*
  Docker, Container, Kubernetes, Distributed Systems
  - We have implemented a movie booking system organized as a set of three microservices: User, Wallet and Booking each hosting a RESTful APIs to handle HTTP requests. To manage load at runtime , we used Kubernetes and deployed the three microservices as load balanced services.
- **Performance Optimization of Dilated Convolution** *[2023]*
  Perf, CUDA , pthread , Computer Architecture
  - We optimized the given Dilated Convolution code for single thread and multi thread using Loop unrolling, Code motion , SIMD Vectorization , pthread and GPU (CUDA Programming), consists of design considerations of kernel and SIMT.In this we achieved maximum speedup 93.2% in case of optimization using GPU.

Experience

- **Cisco** *May 2024 - June 2024*
  Software - Intern Bangalore
  - Developed APIs for the ACIA , allowing seamless integration with other modules. Debugged Python test scripts, identifying and resolving key performance issues .
- **Capgemini** *Aug 2021 - July 2022*
  Senior Analyst Mumbai
  - Worked on various cybersecurity aspects including vulnerability analysis in virtual machines and virtual device drivers,identifying and mitigating vulnerabilities of virtualized environments.

Technical Skills and Interests

Languages: C, C++, Python

Tools/Technologies: Perf, Linux, Docker, Kubernetes, PyTorch, Git , CUDA , LLVM

Coursework: High Performance Computer Architecture, Systems for Machine Learning ,Principles of Distributed Software, Operating Systems, Design and Analysis of Algorithms, Compiler Design

Achievements

- Secured All India Rank 90 in GATE 2023 (Computer Science).
- Samsung Fellowship Awardee (2017-2021) and Jay Pullur Mallika Fellowship Awardee .(2023-2025)
- Winner in inter-NIT chess tournament and badminton events.