Sakshi Sindhwal

M.Tech CSE
Indian Institute of Science, Bangalore

+91-9808109382
sakshi.dakshana16@gmail.com
ssakshi@iisc.ac.in
github.com/ssindhwa/Projects
linkedin.com/in/sakshi-sindhwal-507b39190/

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

Projects

- **LLM Inference Optimization on CPUs** *[2024-25]*
  Python, PyTorch, M.Tech Project at High Performance Computing Lab.
  – Developed a hybrid framework that mitigates the KV cache copy overhead of Attention Sink and memory overhead of Balancing Memory and Compute, improving throughput upto 2.5x for streaming LLM inference on CPUs.
  – Implemented mixed-precision KV cache management, storing important tokens in FP32 while keeping other in INT8 resulted into throughput benefit of 4x compared to baseline in prefill phase that is more compute bound.
  – Exploring KV cache layout in a decoder-only transformer by transitioning tensor format from [B, H, N, D] to [N, B, H, D], improving data locality and minimizing memory access latency on CPU. Reimplementing the core inference kernel to support hardware-efficient patterns and enhancing overall inference performance.
- **Microservice Implementation for Booking System using Spring** *[2024]*
  Docker, Container, Kubernetes , Java, Distributed Systems
  – 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, used Kubernetes and deployed the three microservices as load balanced services.
  – Reimplemented the Booking microservice using Akka, where we designed a Show actor to be created for each show. This Show actor is responsible for maintaining information about all bookings made for that show.
- **Optimized Performance of Dilated Convolution** *[2023]*
  C++, pthreads, Perf, SIMD, CUDA, High Performance Computer Architecture
  – Applied advanced optimization techniques such as loop unrolling, elimination of redundant computations, strength reduction, and SIMD to enhance the performance of the dilated convolution algorithm.
  – Developed and optimized a multi-threaded version using pthreads of the dilated convolution algorithm, leveraging parallel processing for both CPUs and GPUs.
- **System call Sandbox for an Application** *[2024]*
  LLVM, C , Python, eBPF, Computer Systems Security
  – Analyzed source-level C programs and emitted a policy of acceptable library calls generated by the program.
  – Developed an LLVM-based tool that takes a C program as input and generates a library call graph.
  – Extended the project to detect and terminate processes that invoke library calls outside of the predefined sequence policy, flagging them as potential malicious activity to enhance runtime security and integrity.

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, identifying and mitigating vulnerabilities of virtualized environments.

Technical Skills and Interests

Programming Languages: C, C++, Python, Java

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

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

Achievements

- Secured All India Rank 90 in GATE 2023 (Computer Science).