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]*
  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 optimizations by altering the conventional tensor format to identify a more hardware-friendly structure. Investigating different formats to improve memory access latency for CPUs.
- **Memory Checkpointing feature using eBPF** *[2024]*
  C, Python, eBPF, Operating Systems
  – Designed and executed a system to capture and restore process memory states using eBPF.
  – Created tracepoint handlers for system calls to intercept and manage memory operations, enhancing process memory tracking and logging.
  – Addressed challenges such as excluding stack VMAs during checkpointing and ensured efficient memory write-back using eBPF helpers, resulting in robust state restoration capabilities.
  – Devised and integrated data structures for efficient data management between user and kernel space in Linux OS.
- **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.
- **Optimizing Performance of Dilated Convolution** *[2023]*
  C++, pthreads, Perf, SIMD, CUDA, 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.

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*
  Software Engineer 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

Tools/Technologies: Perf, Linux, Git , LLVM , eBPF

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).