Rahul Sharma

(408) 623-2823 | rahuldsharma17@berkeley.edu | linkedin.com/in/rahul-sharma17 | San Jose, CA

Education

University of California, Berkeley Aug 2021 – May 2025
Bachelors in Computer Science | ***GPA**: 4.0/4.0* *Berkeley, CA*
Courses: Machine Learning, Operating Systems, Efficient Algorithms, Advanced LLM Agents, Artificial Intelligence, Computer Security (TA), Data Structures, Convex Optimization, Computer Architecture, Principles of Data Science

Experience

NVIDIA May 2024 – Aug 2024
Deep Learning Algorithms Intern - Model Optimization *Santa Clara, CA*
- Developed a platform leveraging NVIDIA NeMo/TensorRT-LLM, Docker, and Slurm to deploy distributed LLM training/inference workloads across internal GPU clusters, supporting runs on multiple GPUs and nodes
- Benchmarked LLM (LLaMa3, Mistral, etc.) throughput performance (tok/sec) across DGX servers and lower-cost GPU architectures (A100, L40S, H100) to identify cost-optimized deep learning datacenter configurations
- Optimized model execution on low-cost servers via model parallelism, hyperparameter tuning, FP8 quantization, and mixed precision, achieving up to 15% throughput increase on various workloads
- Built a pipeline using Nsight Systems and SQLite to track NVLink bandwidth across distributed LLM workloads, analyzing GPU communication bottlenecks to further optimize performance

Berkeley Artificial Intelligence Research (BAIR) Aug 2023 – May 2024
Researcher - Computational Imaging Lab *Berkeley, CA*
- Utilized PyTorch to contribute towards the development of a UNet-based model incorporating multi-attention and sinusoidal positional embeddings for enhanced image denoising and reconstruction
- Investigated the impact of applying and learning a point spread function on image reconstruction quality versus traditional downsampling with Gaussian noise, resulting in a 5% accuracy improvement with the new approach
- Worked on building a scalable video-caption generation pipeline and finetuning text-to-video models

Amazon May 2023 – Aug 2023
Software Engineering Intern - Alexa Devices Lab126 *Sunnyvale, CA*
- Developed a Python-based framework to conduct real-time performance monitoring/testing for Amazon Echo device cameras, validating KPIs like resolution, frame rate, and image capturing latency
- Implemented scalable backend services for storing/analyzing the reported camera performance data, utilizing PostgreSQL for metadata and event logging and Amazon S3 for storing captured images and videos
- Collaborated with the computer vision team to enhance motion detection capabilities, enabling the automatic adjustment of camera features based on detected movement

GreenOps May 2022 – Aug 2022
Software Engineering Intern *San Jose, CA*
- Built on top of a Kubernetes infrastructure to deploy ArgoCD applications within the GreenOps environment, conducting stress and load testing to measure overall system reliability
- Used Prometheus/Grafana to set up metrics endpoints, scrape CPU/memory usage logs, and create visualizations

Projects

PintOS — *C, x86* Jan 2024 – May 2024
- Created a custom operating system for the x86 architecture, implementing features such as kernel threads, a buffer caching file system, floating point operations, and virtual memory
- Enhanced processing speeds by over 40% via synchronization, priority scheduling algorithms, resource allocation strategies, and efficient memory management techniques

Secure File Storage System — *Golang* Mar 2024 – May 2024
- Designed a secure end-to-end file storage system using Golang and cryptography concepts to ensure data integrity and confidentiality (symmetric key encryption, message authentication codes, hashing, etc.)
- Implemented account management and file operations: create, read, overwrite, append, and file sharing

Convolution-Optimized Video Processing — *C, OpenMP* Oct 2023 – Dec 2023
- Optimized 2D convolution operations for video processing, leveraging SIMD instructions, OpenMP directives, cache/algorithmic optimizations to achieve a 12x performance speedup
- Constructed an Open MPI coordinator for parallel task execution, improving efficiency through multiprocessing

Technical Skills

Languages: Python, Java, C, C++, Rust, CUDA, Golang, Javascript, x86, RISC-V, HTML/CSS, Bash/Shell, R
Machine Learning/Data: PyTorch, TensorFlow, NumPy, Pandas, SQL, Scikit, OpenCV, TensorRT, NeMo
Technologies: Linux, Docker, Git, AWS, GCP, Kubernetes, Slurm, React, Node, Promethus, Grafana, PostgreSQL