Shreyashri Biswas

shreyabw834@gmail.com | shreyab8.github.io | linkedin.com/in/shreyab8/ | +1(412) 909-9766

EDUCATION

Carnegie Mellon University, Pittsburgh, PA Dec 2023
Master of Science in Electrical and Computer Engineering
Courses: *Intro to Deep Learning(11785), Modern Computer Architecture(18740), On Device ML(11767), How to Write Fast code: GPU Parallel Prog.(18646), Optimization(18660), Visual Learning and Recognition(16824), Large Language Models(11667), Deep learning systems(10714)*
TA: *Parallel Computer Architecture and Programming* (15-618)

SRM Institute of Science and Technology, Chennai, India (CGPA 9.25/10 , **Graduated with Honors**) May 2022
Bachelor of Technology in Electronics and Communication Engineering with Minors in Computer Science and Engineering
Highlights*: Best Bachelor Thesis Project, 2nd at Github Hackathon, 10th at Bosch Hackathon, Won SIIC projects, Published a sponsored patent*

SKILLS

Languages/Framework: C++, Python, PyTorch, Tensorflow, Jax
HPC/ MLSys: Edge AI, Deep Learning, Machine Learning, SIMD, CUDA, OpenMP, OpenMPI, OpenGL, VitisAI, ARMNN, TVM, ISPC, MLIR
Computer Architectures: CPU (x86, ARM), GPU (NVIDIA), Qualcomm DSP, CGRA (AMD AI Engine)

EXPERIENCE

Software Developer 2, AI Models Team, MLSE Feb 2024 - Present
AMD, Austin, Texas, US
- **ROCm GEMM Tuning & Profiling**: Tuned GEMM operations (hipBLASLT) for Llama-7b, reducing inference time by ~31%. Optimized Residual kernels and integrated performance profiling, boosting end-to-end model efficiency.
- **Performance Optimization for Llama.cpp**: Achieved measurable speedups via wavefront-based parallelism, refined memory access patterns, minimized synchronization overhead and rooftop analyses for identifying high-impact kernels and pinpointing non-performant operations
- **Cross-GPU Vision Model Profiling**: Profiled and improved performance on MI300 vs H100 for vision models, closing the gap from ~61% to ~76%, and generated actionable insights for GEMM and Triton kernel optimization.
- **GPU Acceleration & CuPy Contributions**: Drove CuPy's ROCm platform enablement by hipifying NVTX, resolving failing unit tests, and merging major ROCm branches, enhancing GPU support and code reliability.

Applied Machine Learning System Architecture Engineering Intern (Radeon Technology group) May 2023 - August 2023
AMD, Santa Clara, California, US
- Explored the Pareto optimal of model size and inference latency on AMD's AI accelerator on edge leveraging Quantization techniques
- Employed Quantization Aware Training of VitisAI on an Autoencoder with 10x size reduction & reduced latency by 23% within 2% acc drop
- Identified and fixed over 4 issues in Vitis AI's production compiler stack related to upsampling operation within a time-critical deadline of 4 week

Research Intern – Accident Research Team under Advanced Autonomous Driver Systems (ADAS) Nov 2021 - Feb 2022
Bosch, Bangalore, India
- Designed an end-to-end system with 88% accuracy to predict accident-prone junction areas on Indian roads.
- Translated road accident statistics into satellite imagery-based classifications, pinpointing high-risk junctions. Utilized Pandas for data exploration, OpenCV for image processing, and Gmaps APIs for real-time satellite data.

RESEARCH EXPERIENCE

Research Assistant in Catalyst Lab with Prof. Zhihao Jia (Jan 2023- Present)
- Working on SpecInfer: **Accelerating LLM inference** through speculative inference and token tree verification to predict potential outputs from LLM using smaller models and **tree-based parallel decoding**, allowing simultaneous verification of multiple predicted responses. [ref]
- Engineered a **compiler system for deep learning graph optimization** and op-fusion targeting mobile backends with heterogeneous computer architecture. Implemented a kernel mapping strategy leveraging **TVM and TASO**, enhancing performance

Research Assistant in Speed Lab with Prof. Tze Meng
- Designed **high performant kernels CPU and GPU** for Computer Vision blurring algorithms
- Leveraged performance engineering techniques such as **instruction pipelining, tiling, memory optimization, vectorizing and occupancy optimization** to maximize hardware utilization achieving 17% speedup benchmarked against **OpenCV C++ kernels**

SELECTED ACADEMIC PROJECTS

Accelerating Max-Pool sub-graph in YOLOV4
- Achieved **53% speed-up** with custom designed **CUDA kerne**l leveraging **kernel fusion, padding elimination & memory footprint reduction**
- Addressed performance bottlenecks (**unnecessary computational cycles, excessive memory allocation, and limited fusion opportunities**) through strategic kernel launch configurations, and adaptability to accommodate larger filter chains and non-singular stride patterns

Swin-Vision Transformer for edge device
- **Linearized self-attention** in Transformers for Jetson Nano achieving **14-20% memory and 2-3% latency savings**
- Leveraging (q, k, v) low-rank structures, applied LoRa approx. to the self-attention weight matrices and harnessed SVD in attention block for efficient matrix operations allowed optimizing vision transformers saving, **18% GFLOPS** with 2% accuracy drop

Memory Fairness Protocol | CMU 18740 Modern Computer Architecture
- Developed both custom cache mechanism based on the min-max fairness algorithm and a custom scheduler using starvation queue policy idea to give **weighted speedup of 1.98**

Stress testing and Improved performance of Asymmetric cores of SoC (Snapdragon 888) | CMU 18740 Modern Computer Architecture
- Optimized performance-power trade-offs for a stress-tested SoC by implementing low instruction-level parallelism (ILP) and multithreaded programming, evaluated using Dhrystone and Linpack benchmarks.
- Achieved a **6% misprediction rate**, **10% L1 cache miss rate**, and reduced **L2 cache bottlenecking by 45%**, significantly improving overall system efficiency