Srinivasan Subramaniyan

The Ohio State University (US)

+1 740 274 2814 | Webpage| LinkedIn| Google Scholar| subramaniyan.4@osu.edu

Summary

Ph.D. candidate in Electrical and Computer Engineering at The Ohio State University, focusing on GPU scheduling, data center efficiency, and LLM/AI systems. Awarded Best Paper honors at EMSOFT 2025 and VLSID 2022. Seeking internships in computer engineering, software systems, or high-performance computing.

Education

M.S. + Ph.D. (Computer Engineering) The Ohio State University 2021-present

Technical Skills

- Programming Languages: x86/ARM/RISC-V Assembly, C/C++, Python, Bash Scripting
- Hardware Design & Verification: Verilog, SystemVerilog, FPGA/SoC Design, Hardware Simulation & Debugging
- Optimization & Modeling Tools: Gurobi, PuLP, Simulink, Performance Profilers (gprof, perf, NVProf, Nsight)
- Simulation and Design Software: MATLAB, Vitis Design Suite, Android Studio, Verilator, Gem5, ModelSim
- Parallel Computing: OpenCL, CUDA, OpenMP, HIP, MPI
- Development Tools: Git, Linux Kernel Modules, Docker, Kubernetes, vLLM

Industrial Experience

AMD **May 2022 – Aug 2022**

Research Intern, ***Supervisors:*** *Derrick Aguren and Pete Ehrett* *Austin, US*

- Worked with the RAD (Research and Development) team on GPU performance analysis and kernel optimization.
- Investigated bottlenecks in existing BLAS libraries, motivating architectural improvements for future AMD GPU designs.
- Designed and implemented an optimized `MTTKRP` HIP kernel for AMD GPUs to accelerate sparse tensor operations.

2. ***Centre for Heterogeneous and Intelligent Processing Systems*** **Jan 2019 – Aug 2021**

(Junior Research Fellow) *Bangalore, IND*

- Conducted design space exploration for NB-LDPC codes on FPGAs (Published: SIPS '20, IEEE Design & Test '22).
- Developed accelerators for sparse matrix multiplication (Published: VLSID '22).

Research Projects

1. Real-Time GPU Scheduling

(Guide: Prof. Xiaorui Wang)

- Designing a two-tier feedback control framework for spatially shared GPUs in real-time systems (In Submission 25).
- Proposed **FC-GPU**, the first feedback-control GPU scheduling framework for real-time systems that uses a MIMO controller to adapt task rates dynamically, reducing deadline misses by 2% on RTX 3090 and MI100 platforms (EMSOFT '25). Best Paper Candidate, EMSOFT 2025.

2. Latency-Controlled and Cost-Efficient GPU Scheduling for AI Workloads

(Guide: Prof. Xiaorui Wang)

- Designed **CapLLM**, a power-capping framework for LLM-serving data centers that dynamically manages GPU power to minimize performance violations under energy constraints (In Submission '25).
- Developed a correlation-aware scheduler that consolidates negatively correlated ML workloads on shared GPUs using DVFS control, reducing *OpEx* through improved utilization (In Submission '25).
- Developed **SEEB-GPU**, an edge inference framework that jointly optimizes batching, early exits, and GPU partitioning to reduce latency by up to **15×** while ensuring SLA compliance (SEC '25).
- Proposed **CorrGPU**, a correlation-aware GPU scheduler that dynamically pairs complementary workloads to reduce contention and lower *CapEx* by **20.88%** in large-scale ML traces (IPCCC '25). Best Paper Runner-up, IPCCC 2025.
- Implemented **CapGPU**, a coordinated CPU–GPU power-capping strategy that improves inference throughput by 8–20% while maintaining latency SLOs under power constraints (ICPP '25).

- Built **GPUColo**, a co-location framework that enables training and inference workloads to share GPUs, saving up to **74.9%** of GPUs and reducing *CapEx* with strict SLO compliance (ICDCS '24).

Achievements/Awards

- ***IPCCC Best Paper Runner-Up Award:*** 44th IEEE International Performance Computing and Communications Conference (IPCCC 2025).
- ***EMSOFT Outstanding Paper Award:*** International Conference on Embedded Software (EMSOFT 2025).
- ***EMSOFT Travel Grant Award*** (2025).
- ***BurnLin Travel Grant Award*** (2023, 2024, 2025).
- ***A.K. Choudhary Best Paper Award:*** 35th International Conference on VLSI Design and the 21st International Conference on Embedded Systems (VLSID 2022).
- ***Amrita Scholarship*** awarded during undergraduate studies at Amrita Vishwa Vidyapeetham.

Selected Publications

- Srinivasan Subramaniyan and Xiaorui Wang. "FC-GPU: Feedback Control GPU Scheduling for Real-time Embedded Systems." *Embedded Systems Week – International Conference on Embedded Software (EMSOFT)*, 2025. Outstanding Paper Award.
- Srinivasan Subramaniyan and Xiaorui Wang. "Exploiting ML Task Correlation in the Minimization of Capital Expense for GPU Data Centers." In *Proceedings of the 2025 IEEE International Performance, Computing, and Communications Conference (IPCCC)*. IEEE, 2025. Best Paper Runner-up award
- Yuan Ma, Srinivasan Subramaniyan, and Xiaorui Wang. "Power Capping of GPU Servers for Machine Learning Inference Optimization" *54th International Conference on Parallel Processing (ICPP)*, 2025.
- Chen, Guoyu, Srinivasan Subramaniyan, and Xiaorui Wang. "Latency-Guaranteed Co-Location of Inference and Training for Reducing Data Center Expenses" *IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, 2024.

Course Work

- Computer Architecture, Embedded Systems, Operating Systems, Hardware Architecture Techniques, Parallel Computing, Algorithms, Reinforcement Learning & Machine Learning, Parallel and Distributed Systems, High-Performance Computing (HPC), FPGA/SoC Design and Performance Modeling & Optimization.

Positions of Responsibility

Treasurer

1. IEEE Graduate Student Body (GSB), Jan 2025 – Present *The Ohio State University*

- Oversee financial accounts, budgeting, and allocation of funds to ensure responsible management of IEEE GSB resources.
- Planned, hosted, and tracked funding for technical seminars, networking mixers, and professional development events for graduate students.
- Reinstated the organization from inactive to active status through compliance work and renewed student engagement.