

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 full-time 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
- Parallel Computing: OpenCL, CUDA, OpenMP, HIP, MPI, Rocm
- Development Tools: Git, Linux Kernel Modules, Docker, Kubernetes, vLLM

Industrial Experience

AMD

Research Intern, Supervisors: Nuwan Jayasena, Mahmoud Khairy Abdallah

May 2026 -

Santa Clara, US

- Designed NUMA-aware optimizations for LLM inference workloads on multi-GPU platforms, improving memory locality and reducing cross-socket communication overhead.

AMD

Research Intern, Supervisors: Derrick Aguren and Pete Ehrett

May 2022 – Aug 2022

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)

- Proposed and designed **JTPC**, a joint timeliness and power control framework based on an LQR-based MIMO controller for performance and power optimization in real-time embedded GPU systems ([Under submission 2026](#)).
- Proposed and designed **PowerGPU**, a two-layer coordinated response-time ratio (RTR) control framework combining GPU DVFS and rate adaptation for power-efficient real-time embedded systems, achieving 29.4% lower power than prior RTR approaches ([Under submission, 2026](#)).
- Proposed and designed **DySM**, a closed-loop GPU scheduling framework leveraging dynamic SM scaling and MIMO feedback control to guarantee response-time targets in spatially shared real-time systems ([ECRTS' 2026](#)).
- 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)

- o Designed **LLMcap**, a power-capping framework for LLM-serving data centers that dynamically manages GPU power to minimize performance violations under energy constraints ([In Submission '26](#)).
- o Developed **CATS** a correlation-aware scheduler that consolidates negatively correlated ML workloads on shared GPUs using DVFS control, reducing *OpEx* through improved utilization ([ICS '26](#)).
- o Developed **SEEB-GPU**, an edge inference framework that jointly optimizes batching, early exits, and GPU partitioning to reduce latency by up to **15x** while ensuring SLA compliance ([SEC '25](#)).
- o 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**.
- o 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](#)).
- o 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

- o **Hayes Graduate Research Forum:** 3rd Place, The Ohio State University (2026).
- o **College of Engineering (CoE) Winner:** Three Minute Thesis (3MT) Competition, The Ohio State University (2026).
- o **IPCCC Best Paper Runner-Up Award:** 44th IEEE International Performance Computing and Communications Conference (IPCCC 2025).
- o **EMSOFT Outstanding Paper Award:** International Conference on Embedded Software (EMSOFT 2025).
- o **EMSOFT Travel Grant Award** (2025).
- o **BurnLin Travel Grant Award** (2023, 2024, 2025).
- o **A.K. Choudhary Best Paper Award:** 35th International Conference on VLSI Design and the 21st International Conference on Embedded Systems (VLSID 2022).
- o **Amrita Scholarship** awarded during undergraduate studies at Amrita Vishwa Vidyapeetham.

Selected Publications

- o Srinivasan Subramaniyan* and Xiaorui Wang. "DySM: Dynamic Scaling of GPU Streaming Multiprocessor in Spatially Shared Real-Time Embedded GPU Systems." *38th Euromicro Conference on Real-Time Systems (ECRTS)*, 2026.
- o Srinivasan Subramaniyan* and Xiaorui Wang. "CATS: Correlation-aware Task Scheduling for GPU Power Optimization in AI Data Centers." *40th ACM International Conference on Supercomputing (ICS)*, 2026.
- o 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**.
- o Srinivasan Subramaniyan*, Rudra Joshi, Xiaorui Wang, and Marco Brocanelli*. "SEEB-GPU: Early-Exit Aware Scheduling and Batching for Edge GPU Inference." *10th ACM/IEEE Symposium on Edge Computing (SEC)*, 2025.
- o 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.
- o 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

- o 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
- o 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.