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, specializing in GPU scheduling, computer architecture, and edge/cloud systems. Author of multiple award-winning papers, including the Outstanding Paper Award at EMSOFT 2025 and Best Paper Award at VLSID 2022. Actively seeking opportunities in computer architecture, high-performance computing, and systems research.

Education

PhD (Computer Engineering) The Ohio State University 2021-

Technical Skills

- Programming Languages: x86/ARM/RISC-V Assembly, C/C++, Python, Bash Scripting
- Hardware Design & Verification: Verilog, SystemVerilog, VHDL, 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, ROCm Stack, Android Studio, Verilator, Gem5, ModelSim
- Parallel Computing: OpenCL, CUDA, OpenMP, HIP, MPI
- Development Tools: Git, Linux Kernel Modules, Docker, Kubernetes

Research Projects

1. Integrated Feedback Control Framework for Real-Time GPU Scheduling in Autonomous Systems
(Guide: Prof. Xiaorui Wang)

- Proposed the first feedback control-based GPU scheduling (**FC-GPU**) framework for real-time systems, using a MIMO controller to dynamically adapt task rates (EMSOFT '25).
- Designing a two-tier feedback control framework for spatially shared GPUs in real-time systems (In Submission 25).

2. Latency-Controlled Reduction of Data Center Expenses for AI Workloads
(Guide: Prof. Xiaorui Wang)

- Developed a correlation-aware scheduling algorithm to consolidate negatively correlated ML workloads on shared GPUs, integrating DVFS to reduce OpEx (In Submission '25).
- Designed a correlation-aware GPU scheduling algorithm (**CorrGPU**) to minimize CapEx in data centers for ML workloads (IPCCC '25).
- Implemented power-capping strategies (**CapGPU**) for ML inference servers to reduce energy consumption while meeting latency SLOs (ICPP '25).
- Built a co-location framework (**GPUColo**) to consolidate ML inference and training workloads on the same GPUs, ensuring SLOs are met while reducing *CapEx* (ICDCS '24).

3. Power-Aware and Resource-Efficient Edge Computing
(Guide: Prof. Xiaorui Wang and Prof. Marco Brocanelli)

- Developed SEEB-GPU, an inference framework for edge GPUs that integrates deadline-aware batching, confidence-based early exits, and GPU spatial isolation to achieve up to **7× latency reduction** while ensuring SLA compliance (In Submission '25).

Industrial Experience

1. *AMD* **May 2022 - Aug 2022**
Research Intern *Austin, US*

- Optimized the scheduling of GP-GPU kernels to accelerate graph-based applications, enhancing performance and efficiency.
- Discovered optimization strategies for matrix multiplications involving tall and wide matrices, significantly boosting overall performance. (Published: IPDPSW '23).

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).
- Appointed as a Visiting Research Fellow at the Instituto de Telecomunicações, University of Coimbra, from March 2021 to June 30, 2021.

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.

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.
- 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.

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.

Achievements/Awards

- ***EMSOFT Outstanding Paper Award:*** International Conference on Embedded Software (EMSOFT 2026).
- ***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.