Srinivasan Subramaniyan

The Ohio State University (US)

525 Harley Drive APT 6, Columbus - 43202, Ohio, US
☎ +1 740 274 2814 • ✉ subramaniyan.4@osu.edu
www.linkedin.com/in/srinivasan-subramaniyan22 • Webpage

Summary

I am a committed Ph.D. student in Computer Engineering with four years of research experience. I am actively pursuing an internship opportunity that will allow me to apply my expertise in Computer Systems and Engineering.

Education

PhD (Computer Engineering)	The Ohio State University	**3.83/4**	2022-
M.S (Computer Engineering)	The Ohio State University	**3.83/4**	2021-2024

Technical Skills

- Programming Languages: x86/ARM/RISC-V Assembly, C/C++, Python, Bash Scripting
- Parallel Computing: OpenCL, CUDA, OpenMP, HIP
- Simulation and Design Software: Matlab, Vitis Design Tools, ROCm Stack, Android Studio, Verilator
- Optimization Solvers: Gurobi, PuLP

Research Projects

1. *An Integrated Feedback Real-time Scheduling Framework for GPU-based Autonomous Systems* **Jan 2024 -**
(Guide: Prof Xiaorui Wang)

- Developed a MIMO-controller-based design for dynamic task rate adaptation in GPUs used in Real-time Systems (In Submission: ASPLOS 24).
- Extensive hardware testbed results on an Nvidia GPU demonstrated that FC-GPU provided 30% fewer deadline misses for GPU tasks, even with significant runtime execution time increases.

2. *Enabling Latency-guaranteed Co-location of Inference and Training for Reducing Data Center Expenses* **Nov 2022 -**
(Guide: Prof. Xiaorui Wang)

- Two-Tier Design: The outer loop uses MPS to allocate GPU threads dynamically, ensuring inference meets SLOs despite MPS overheads. The inner loop implements periodic training sleep to rapidly free GPU resources for responsive inference latency control.
- Performance Impact: Precisely controls latency while saving 72.66% of GPUs, reducing capital costs based on a large-scale data center trace for a 57 day old trace.(ICDCS 2024).

3. *Correlation Aware scheduling of Machine Learning Workloads in Datacenters* **Aug 2023 -**
(Guide: Prof. Xiaorui Wang)

- Designed a novel task scheduling algorithm to consolidate negatively correlated ML tasks onto the same GPUs, reducing job completion times and reduce CapEx (In Submission: INFOCOM 24) (Submitting to: IPDPS 24).
- Our simulation results on real-world Machine Learning traces also show that CorrGPU outperforms several state-of-the-art solutions by having 55.8% less capital expense

Industrial Experience

1. *AMD* **May 22 - Aug 22**
(Research Intern) *Austin,US*

- Worked on scheduling GP-GPU kernels for graph applications.
- Identified and analyzed performance bottlenecks for graph applications on GPUs.
- Proposed the concept of cache cohorts for efficient kernel scheduling to maximize cache utilization.

2. *Centre for Heterogeneous and Intelligent Processing Systems* **Jan 19 - Aug 21**
(Junior Research Fellow) *Bangalore,IND*

- Conducted design space exploration for NB-LDPC codes on FPGAs.

- Developed accelerators for sparse matrix multiplication.
- Appointed as a Visiting Research Fellow at the Instituto de Telecomunicações, University of Coimbra, from March 21 to June 30th.

Course Work

- Computer Architecture, Embedded Systems, Operating Systems, Hardware Architecture Techniques, Parallel Computing, Reinforcement Learning & Machine Learning

Academic Projects

- Constraint Rectrified Policy Optimization using Reinforcement Learning
- Division of Workload among CPU/GPU
- Design Space Exploration for Dense MTTKRP using OpenMP Offloading
- Transient Execution Attacks: A Comprehensive Survey
- Smart Music Reactive LED Strip
- LNPG: Leaking Neural Network Secrets through Power Side-channel in GPUs

Selected Publications

- 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.
- Srinivasan Subramaniyan, Oscar Ferraz, M. R. Ashuthosh, Santosh Krishna, Guohui Wang, Joseph R. Cavallaro, Vitor Silva, Gabriel Falcao, and Madhura Purnaprajna. "Enabling High-Level Design Strategies for High-Throughput and Low-Power NB-LDPC Decoders" IEEE Design & Test (2022).

Achievements/Awards

- ***BinLin Travel Grant Award*** *(2023 & 2024).*
- ***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.