Sohaib Ahmad

sohaib@cs.umass.edu | sohaibahmad759.github.io | linkedin.com/in/sahmad | Google Scholar

Education

University of Massachusetts Amherst — *Amherst, MA*

Ph.D., Computer Science — *Dec 2019 - Nov 2024 (expected)*

- **Advisors:** Dr. Ramesh Sitaraman, Dr. Hui Guan
- **Thesis:** Adaptive Scheduling and Resource Management for ML Model Serving
- **Interests:** Systems for Machine Learning, Model Serving, Scheduling and Resource Management

University of Massachusetts Amherst — *Amherst, MA*

MS, Computer Science — *Sep 2017 - Dec 2019*

- **Thesis:** Supervised Learning Techniques for Optimizing Energy Usage in Cloud Datacenters
- **Courses:** Systems for ML and ML for Systems, Distributed and Operating Systems, Reinforcement Learning, Machine Learning

Lahore University of Management Sciences — *Lahore, Pakistan*

BS, Computer Science, GPA: 3.92 — *Aug 2013 - May 2017*

- Ranked 1st in the Computer Science class of 2017 (CGPA)
- Dean's honor list for excellent academic performance in all years
- **Courses:** Distributed Systems, Computer Networks, Topics in Computer Networks

Work Experience

Meta (Facebook) — *Bellevue, WA*

Software Engineering Intern — *May 2022 - Aug 2022*

- **Team:** Serverless Computing
- Implemented a new framework to add and manage elastic workers to the pool of serverless compute resources and throttle them based on different SLA requirements of applications
- **Skills:** C++, Python, gTest, Thrift (gRPC equivalent)

Nokia Bell Labs — *Murray Hill, NJ*

Research Intern — *Jun 2021 - Aug 2021*

- Designed a novel approach to handle workload spikes for resource-constrained clusters serving ML models by scaling the accuracy of served requests in response to changing load
- Formulated a reinforcement learning-based mechanism to learn adaptable scheduling policies that reduced SLA violations by up to 46% compared to the existing approach
- **Skills:** Python, Discrete Event Simulation, Reinforcement Learning, OpenAI Gym

Nokia Bell Labs — *Murray Hill, NJ*

Research Intern — *Jun 2020 - Aug 2020*

- Designed a scheduler to optimize the placement of machine learning training jobs in a geographically distributed cloud over multiple continents, based on individual resource and data constraints of training jobs for the Bell Labs internal model training system
- Reduced the makespan of training jobs by up to 52% under normal load and up to 26% under high load conditions compared to the existing model training system

Selected Publications

Loki: A System for Serving ML Inference Pipelines with Hardware and Accuracy Scaling — *HPDC 2024*

Sohaib Ahmad, Hui Guan, Ramesh K. Sitaraman — *Pisa, Italy*

Proteus: A High-Throughput Inference-Serving System with Accuracy Scaling — *ASPLOS 2024*

Sohaib Ahmad, Hui Guan, Brian D. Friedman, Thomas Williams, Ramesh K. Sitaraman, Thomas Woo — *San Diego, CA*

AggFirstJoin: Optimizing Geo-Distributed Joins using Aggregation-Based Transformations — *CCGrid 2023*

Dhruv Kumar, Sohaib Ahmad, Abhishek Chandra, Ramesh K. Sitaraman — *Bangalore, India*

Best Paper Award

Learning from Optimal: Energy Procurement Strategies for Data Centers — *e-Energy 2019*

Sohaib Ahmad, Arielle Rosenthal, Mohammad H. Hajiesmaili, Ramesh K. Sitaraman — *Phoenix, AZ*

Inside the Walled Garden: Deconstructing Facebook's Free Basics Program — *SIGCOMM 2018*

Rijurekha Sen, Sohaib Ahmad, Amreesh Phokeer, Zaid Farooq, Ihsan A. Qazi, David Choffnes, Krishna P. Gummadi — *Budapest, Hungary*

Best Paper Award, SIGCOMM CCR

Awards & Service

2023 **Dissertation Writing Fellowship**, College of Information and Computer Sciences, UMass Amherst
2023 **Best Paper Award**, ACM/IEEE CCGrid
2023 **Graduate Student Representative for Faculty Hiring**, College of Information and Computer Sciences, UMass Amherst
2022 **Shadow Program Committee Member**, ACM EuroSys
2022 **Panelist**, CS Research Night, UMass Amherst
2018 **Best Paper Award**, ACM SIGCOMM CCR
2017 **Krithi Ramamritham Scholarship for Outstanding Student in Systems Research**, UMass Amherst

Skills

Programming Python, C/C++, Java, JavaScript, MATLAB, SQL
Frameworks PyTorch, Keras, gRPC, Gurobi, RESTful APIs, FastAPI, Flask, Google Test (gTest)
Tools Docker, Git, Jupyter, LaTeX, Bash/Shell