Suhas Jayaram Subramanya

@ suhasj@cs.cmu.edu | github.com/suhasjs | suhasjs.github.io

Education

Carnegie Mellon University — Pittsburgh, USA
Doctor of Philosophy (Ph.D) in Computer Science — *2019 – 2025*

Indian Institute of Technology Madras — Chennai, India
Bachelor of Technology (B.Tech) in Computer Science and Engineering — *2013 – 2017*

Research Interests

My research focuses on the intersection of Systems and Machine Learning. I am interested in optimizing large-scale infrastructure to enable efficient machine learning. My research spans two complementary fronts – making better use of existing resources to improve responsiveness and throughputs (cluster scheduling, DL training) and developing new hardware-software co-designs (ANNS, Larger-than-memory computing) to enable new capabilities using recent developments in hardware.
Topics: Cluster Scheduling, Deep Learning Infrastructure, Approximate Nearest Neighbor Search (ANNS), Larger-than-memory Computing

Work Experience

Project Singularity, Microsoft — Seattle, USA
Research Intern — *Jun – Aug '21*

- Analyzed cluster job traces to create representative workloads that can be used to guide exploration of scheduling algorithms for AI workloads in simulation.

Project Turing, Microsoft — Seattle, USA
Research Intern — *May – Aug '20*

- Designed and developed *FreshDiskANN*, an ANNS system capable of concurrent query, insert and delete operations on billion-scale vector indices with millisecond-scale latencies using NVMe SSDs.
- FreshDiskANN currently powers semantic search in many large-scale applications used by millions of users across thousands of organizations worldwide: (a) web-search through Microsoft Bing, (b) semantic search through Microsoft Office Enterprise Search, (c) Retrivial-Augmented-Generation and hybrid search through Azure CosmosDB for NoSQL, and (d) custom vector indices through Microsoft Azure AI Search.

Microsoft Research India — Bangalore, India
Research Fellow — *Jul '17– Jul '19*

- Designed and developed cost-effective machine learning systems that power production pipelines for topic-modeling, extreme multi-label learning, and approximate nearest neighbor search.
- Developed DiskANN, an ANNS system that powers the Bing Web Search using NVMe SSDs, delivering hundreds of billions of k-ANNS queries at sub-5ms latencies at 10x lower cost compared to competing solutions.
- Developed BLAS-on-Flash, a larger-than-memory computing framework to support training of large-scale machine learning models on commodity hardware using modern NVMe SSDs.

Google — Bangalore, India
Software Engineering Intern — *May – Jul '16*

- Developed annotation metrics and production pipelines to understand efficacy of personalization re-rankers.

Publications

Efficient resource-scheduling at scale via continuous optimization
Suhas Jayaram Subramanya, Don Kurian Dennis, Virginia Smith, Greg Ganger
Work-in-progress, 2024.

Sia: Heterogeneity-aware, goodput-optimized ML-cluster scheduling
Suhas Jayaram Subramanya, Daiyaan Arfeen, Shouxu Lin, Aurick Qiao, Zhihao Jia, Greg Ganger
29th Symposium on Operating Systems Principles (SOSP), 2023.

FreshDiskANN: A Fast and Accurate Graph-Based ANN Index for Streaming Similarity Search
Aditi Singh, **Suhas Jayaram Subramanya**, Ravishankar Krishaswamy, Harsha Vardhan Simhadri
Arxiv 2105.09613, 2021.

Pollux: Co-adaptive Cluster Scheduling for Goodput-Optimized Deep Learning
Aurick Qiao, Sang Keun Choe, **Suhas Jayaram Subramanya**, Willie Neiswanger, Qirong Ho, Hao Zhang, Gregory R. Ganger, Eric P. Xing
15th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2021.

PACEMAKER: Avoiding HeART attacks in storage clusters with disk-adaptive redundancy
Saurabh Kadekodi, Francisco Maturana, **Suhas Jayaram Subramanya**, Juncheng Yang, K. V. Rashmi, Gregory R. Ganger
14th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2020.

DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node
Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, Ravishankar Krishaswamy, Harsha Vardhan Simhadri
Neural Information Processing Systems (NeurIPS), Vancouver, 2019.

BLAS-on-flash: An Efficient Alternative for Large Scale ML Training and Inference?
Suhas Jayaram Subramanya, Harsha Simhadri, Srajan Garg, Anil Kag, Venkatesh Balasubramanian
16th USENIX Symposium on Networked Systems Design and Implementation (NSDI), Boston, 2019.

Exploration for Multi-task Reinforcement Learning with Deep Generative Models
Sai Praveen Bangaru, **Suhas Jayaram Subramanya**, Balaraman Ravindran
NeurIPS Deep Reinforcement Learning Workshop, Barcelona, 2016.

Patents

Building a graph index and searching a corresponding dataset
Suhas Jayaram Subramanya, Ravishankar Krishaswamy, Harsha Vardhan Simhadri
US Patent App. 16/582,682, 2020.

Open Source Contributions

DiskANN | *github*

- Open source implementation of DiskANN and Vamana algorithms in C++ for both Linux and Windows. DiskANN supports building and serving of SSD-based indices for k-ANNS queries on `uint8, int8,` and `float` datasets.

BLAS-on-flash | *github*

- Open source implementation of BLAS-on-flash framework and runtime in C++ with sample kernels for matrix operations like `gemm` and `csrmm` and utility kernels like `sort` and `map-reduce`.

Teaching Experience

Advanced Cloud Computing — Carnegie Mellon University
Instructors: Prof. Greg Ganger and Prof. Majd Sakr — *Jan – May '24*

- Teaching Assistant and Lead for the Cloud Scheduling project for the class of 105 graduate students.

Advanced Storage Systems — Carnegie Mellon University
Instructors: Prof. Greg Ganger and Prof. George Amvrosiadis — *Aug – Dec '21*

- Head Teaching Assistant responsible for assignments and projects for the class of 102 graduate students.

Introduction to Machine Learning — IIT Madras
Instructor: Prof. Balaraman Ravindran — *Jan – Apr '17*

- Teaching Assistant on a MOOC hosted on NPTEL with over 6000 registered students. Course contents now archived to allow others to take the course at their own pace.