Amey **Agrawal**

CS Ph.D. Student, Georgia Tech

ameya.info @ agrawalamey12@gmail.com github.com/agrawalamey Google Scholar

Education

Present
Aug 2022
Georgia Institute of Technology **Atlanta, USA**
Ph.D., Computer Science. GPA 4.00/4.00.
Advisor: Prof. Alexey Tumanov | Area: Systems for machine learning, LLM inference systems.

Jul 2018
Aug 2014
Birla Institute of Technology and Science Pilani **Pilani, India**
B.E. (Hons.), Computer Science

Experience

Aug 2024
May 2024
Microsoft Azure Systems Research **Redmond, USA**
Research Intern | Mentor: Dr. Esha Choukse
Built systems to serve large language models with multimillion context length requests.

Aug 2023
May 2023
Microsoft Research **Bangalore, India**
Research Intern | Mentors: Dr. Ramchandran Ramjee, Dr. Bhargav Gulavani
Designed efficient inference systems for large language models.

Aug 2022
Jan 2021
Microsoft Research **Bangalore, India**
Research Software Engineer-II | Mentor: Dr. Muthian Sivathanu
Built parts of the elasticity sub-system that leveraged efficient time sharing of GPUs to provide transparent scaling of deep learning training workloads. This work was done as a part of the Singularity project, Microsoft's planet-scale AI infrastructure service.

Nov 2020
Jul 2018
Qubole Inc. **Bangalore, India**
Member of Technical Staff-II | Mentor: Rohit Karlupia
Worked on various applied machine learning and software engineering problems to enhance Qubole's data science platform. Published research in several top-tier venues.

Dec 2017
Jul 2017
Software Engineering Intern | Mentor: Bharath Bhushan
Built core data-plane components for Qubole's Deep Learning clusters based on TensorFlow and Apache Spark.

Publications

Mnemosyne: Efficiently Serving Multi-Million Context Length LLM Inference Requests
Amey Agrawal, Haoran Qiu, Junda Chen, Íñigo Goiri, Chaojie Zhang, Ramachandran Ramjee,
Alexey Tumanov, Esha Choukse
Under Review

Maya: Deep Learning Performance Optimization using Emulated Virtual GPUs
Srihas Yarlagadda*, **Amey Agrawal***, Elton Pinto*, Hakesh Darapaneni, Mitali Meratwal, Pranavi Bajjuri, Shivam Mittal, Srinivas Sridharan, Alexey Tumanov
Under Review

On Evaluating Performance Of LLM Inference Serving Systems
Amey Agrawal, Nitin Kedia, Anmol Agarwal, Jayashree Mohan, Souvik Kundu, Nipun Kwatra,
Ramachandran Ramjee, Alexey Tumanov
Under Review

Inshrinkerator: Compressing Deep Learning Training Checkpoints via Dynamic Quantization [pdf]
Amey Agrawal, Sameer Reddy, Satwik Bhattamishra, Sarath Nookala, Vidushi Vashishth, Kexin Rong, and Alexey Tumanov
Proceedings of 15th ACM Symposium on Cloud Computing, 2024, Redmond **[SoCC'24]**

Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve [pdf][code][video]
Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani,
Alexey Tumanov, Ramachandran Ramjee
Proceedings of 18th USENIX Symposium on Operating Systems Design and Implementation, 2024, Santa Clara **[OSDI'24]**

Vidur: A Large Scale Simulation Framework For LLM Inference [pdf][code][video]
Amey Agrawal, Nitin Kedia, Jayashree Mohan, Ashish Panwar, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee, Alexey Tumanov
Proceedings of 7th Annual Conference on Machine Learning Systems, 2024, Santa Clara [**MLSys'24**]

Sarathi: Efficient LLM Inference by Piggybacking Decodes with Chunked Prefills [pdf]
Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee
Preprint: arXiv:2308.16369 (2023) [**CoRR**]

Sybill: Deep Learning Workload Tuning with Virtual GPUs [poster]
Srihas Yarlagadda*, **Amey Agrawal***, Sarath Nookala, Pranavi Bajjuri, Shivam Mittal, Alexey Tumanov
ACM Symposium on Cloud Computing Poster, 2023 [**SoCC'23**]

Singularity: Planet-Scale, Preemptible, Elastic Scheduling of AI Workloads [pdf]
Singularity Team, Microsoft
Preprint: arXiv:2202.07848 (2022) [**CoRR**]

Logan: A Distributed Online Log Parser [pdf]
Amey Agrawal, Rajat Gupta, and Rohit Karlupia
Proceedings of IEEE International Conference on Data Engineering, 2019, Macau [**ICDE'19**]

Learning Digital Circuits: A Journey Through Weight Invariant Self-Pruning Neural Networks [pdf][code]
Amey Agrawal, and Rohit Karlupia
Sparsity in Neural Networks Workshop 2021; New in ML Workshop, NeurIPS, 2019, Vancouver [**SNN'21**]

Delog: A Privacy Preserving Log Filtering Framework for Online Compute Platforms [pdf] [dataset]
Amey Agrawal, Abhishek Dixit, Namrata Shettar, Darshil Kapadia, Rohit Karlupia, Vikram Agrawal, and Rajat Gupta
Proceedings of IEEE International Conference on Big Data, 2019, Los Angeles [**BigData'19**]

Honours and Awards

Center for Research into Novel Compute Hierarchies (CRNCH) Fellowship, 2023 [🌐]
- For research of automatic hardware-aware optimization of deep learning training workloads.

School of Computer Science Fellowship, 2022
- PhD fellowship from Georgia Institue of Technology.

Teaching and Leadership Roles

Systems for Machine Learning *Head Teaching Assistant | Georgia Institute of Technology* Fall'24
- Conducted series of lectures on GPU architecture and LLM inference.

Covid Central, Yavatmal *Technical Lead* [demo] Apr'21 - Jul'21
- Led a group of four developers to create a platform to manage Covid-19 patients in hospitals. Provided the system to regional hospitals for no-charge.

Introduction to Neural Networks & Fuzzy Logic *Head Teaching Assistant* [assignments] Aug'17 - May'18
- Introduced Python programming assignments along with a new custom-built evaluation platform. Other responsibilities included coordinating the team of seven teaching assistants to conduct labs, designing assignments and helping students with the term project.

Introduction to Machine Learning *Teaching Assistant* Jan'18 - May'18
- Conducted introductory sessions on the scientific Python ecosystem, and organized tests and programming assignments for over 100 students in the class.

Concepticon Initiative *Founding Member & Lead Coordinator* Oct'14 - Feb'15
- Led a team of fifty volunteers to organize career awareness events among high school students. These events witnessed participation of over seventeen hundred students across three cities.