

Swapnil Gandhi

gandhis@stanford.edu • <https://swapnilgandhi.com/>

RESEARCH INTERESTS

I am interested in the system-side problems associated with training, deploying, and operationalizing generative models at scale.

EDUCATION

Ph.D. Candidate in Computer Science, Stanford University, GPA: 4.1/4.0 Sep 2022 – Present

Advisor: Prof. Christos Kozyrakis

Research: Adaptive Software Systems for Reliable, Scalable, and Fair Generative AI

M.Tech. (Research) in Computer Science, Indian Institute of Science (IISc) Aug 2017 – Jan 2020

🏆 NetApp Gold Medal and Best M.Tech. (Research) Thesis (Honorable Mention) for “Distributed Programming Abstraction for Scalable Processing of Temporal Graphs”

B.Tech. in Computer Engineering, Bharati Vidyapeeth Pune Jul 2010 – Jun 2014

🏆 Department Honors and Best Undergraduate Thesis for “Mutation Testing Tool for C Programs”

PUBLICATIONS

[Papers & Posters available [here](#).]

PEER-REVIEWED CONFERENCES

- [7] Ziyi Xu, Zhiqiang Xie, [Swapnil Gandhi](#), Christos Kozyrakis, “FailSafe: High-performance Resilient Serving”, *Under Review*
- [6] [Swapnil Gandhi](#), Christos Kozyrakis, “MoEtion: Efficient and Reliable Sparse Checkpointing for Mixture-of-Experts Models at Scale”, *To Appear In proceedings of the 23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 2026)*, May 2026.
- [5] Athinagoras Skiadopoulos, Mark Zhao, [Swapnil Gandhi](#), Thomas Norrie, Shrijeet Mukherjee, Christos Kozyrakis, “Accelerating Mixture-of-Experts Training with Adaptive Expert Replication”, *To Appear In proceedings of the 23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 2026)*, May 2026.
- [4] [Swapnil Gandhi](#), Mark Zhao, Athinagoras Skiadopoulos, Christos Kozyrakis “ReCycle: Pipeline Adaptation for the Resilient Distributed Training of Large DNNs”, *In proceedings of the 30th ACM SIGOPS Symposium on Operating Systems Principles (SOSP 2024)*, Nov 2024.
🏆 Selected as SRC JUMP 2.0 Best Paper for Quarter 3 2024.
- [3] Anand Iyer, Mingyu Guan, Yinwei Dai, Rui Pan, [Swapnil Gandhi](#), Ravi Netravali “Improving DNN Inference Throughput Using Practical, Per-Input Compute Adaptation”, *In proceedings of the 30th ACM SIGOPS Symposium on Operating Systems Principles (SOSP 2024)*, Nov 2024.
- [2] [Swapnil Gandhi](#), Anand Padmanabha Iyer, “P³: Distributed Deep Graph Learning at Scale”, *In proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2021)*, Jul 2021.
- [1] [Swapnil Gandhi](#), Yogesh Simmhan, “An Interval-centric Model for Distributed Computing over Temporal Graphs”, *In proceedings of the 36th IEEE International Conference on Data Engineering (ICDE 2020)*, Dallas, Texas, April 2020.

PEER-REVIEWED POSTERS

- [1] [Swapnil Gandhi](#), “Wave: A Substrate for Distributed Incremental Graph Processing on Commodity Clusters”, *2nd ACM Student Research Competition (SRC) at 27th Symposium on Operating Systems Principles (SRC- SOSP 2019)*, Ontario, Canada, Oct 2019.
🏆 Received Bronze Medal, Student Research Competition (Graduate Category)

AWARDS & HONORS

Stanford Computer Science Student Service Award 2025	Jun 2025
Selected for the 2025 MLCommons ML and Systems Rising Stars Cohort Part of a global cohort of 38 Ph.D. students in Machine Learning Systems. 🔗	Apr 2025
Joint University Microelectronics Program (JUMP) 2.0 Best Paper For “ReCycle: Pipeline Adaptation for the Resilient Distributed Training of Large DNNs” in Quarter 3 2024. 🔗	Nov 2024
Stanford Computer Science Student Service Award 2024	Jun 2024
Stanford Computer Science Student Service Award 2023	Jun 2023
NetApp Gold Medal for Best M.Tech (Research) Thesis (Honorable Mention), IISc For “Distributed Programming Abstraction for Scalable Processing of Temporal Graphs”. 🔗	Jan 2022
Bronze Medal, 2 nd ACM Student Research Competition (Graduate Category), at SOSP For “Wave: A Substrate for Distributed Incremental Graph Processing on Commodity Clusters”. 🔗	Oct 2019
Won 12 th IEEE International TCSC Scalable Computing (SCALE) Challenge For “Dynamic Scaling of Video Analytics for Wide-area Tracking in Urban Spaces”. 🔗	May 2019
Best Poster Award, 10 th EECS Research Students Symposium, IISc Bangalore For “Distributed Processing Model For Temporal Graphs”. 🔗	Apr 2019
Best Student Research Symposium Poster, IEEE HiPC, Jaipur, India For “Distributed Querying over Compressed Property Graphs”.	Dec 2017
Department Honors, Bharati Vidyapeeth, Pune For outstanding academic performance (Batch 2010 – 2014). 🔗	Jun 2014
TCS Popular Student Project, Bharati Vidyapeeth, Pune For “Mutation Testing Tool for C Programs”, Bachelors dissertation.	May 2014
Best Undergraduate Thesis Award, TRDDC Annual Students Day, Pune For “Mutation Testing Tool for C Programs”, Bachelors dissertation.	Apr 2014

INDUSTRY EXPERIENCE

NVIDIA , Research Intern Internship Mentors: Siva Hari, William J Dally <i>Investigated interactions between inter- and intra-request parallelism in LLM inference; designed runtime mechanisms for dynamic parallelism control to improve fairness</i>	Jun 2025 – Present
Microsoft Research India , Research Fellow Internship Mentor: Anand Iyer <i>Researched techniques for improving system-wide inference goodput for early-exit deep neural network using heterogeneous resources; led to a paper in SOSP 2024</i>	Jul 2021 – Sep 2022
Microsoft Research India , Research Intern Internship Mentor: Anand Iyer <i>Researched how model and data parallelism can be combined with independent graph partitioning for training Graph Neural Networks (GNNs) at scale; led to paper in OSDI 2021</i>	Sep 2020 – Mar 2021
Microsoft Research India , Research Intern Internship Mentors: Karthik Ramachandra, Bhargav Gulavani <i>Investigated and implemented query optimizer modifications to overcome performance regressions in scalar UDF inlined queries; shipped in SQL Server.</i>	Mar 2020 – Aug 2020

SERVICE

Co-Chair, Stanford CS PhD Admit Weekend	2022 – Present
Co-Organizer, Stanford CS Application Assistance Program (SASP)	2022 – Present
Artifact Evaluation Committee (AEC) Member, ACM EuroSys 2023	Aug 2022 – Oct 2022
Shadow PC Committee Member, ACM EuroSys 2022	Oct 2021 – Dec 2021
Artifact Evaluation Committee (AEC) Member, USENIX OSDI 2020	Aug 2020
Artifact Evaluation Committee (AEC) Member, ACM ASPLOS 2020	Dec 2019
Artifact Evaluation Committee (AEC) Member, ACM SOSP 2019	Aug 2019
Treasurer and General Secretary for IISc ACM Student Chapter	Apr 2019 – Mar 2020

**TEACHING
ASSISTANTSHIPS**

CS 349D: Cloud Computing Technology, Stanford

Apr 2024

Graduate Course Assistant for CS 349D. Handled class discussion, homework assignments and helped with class projects (\approx 30 students). 🏆 Received Outstanding CA Award (CS Department).

DS 256: Scalable Systems for Data Science, IISc

Jan 2019

Graduate Teaching Assistant for DS 256. Handled weekly discussion sections, homework assignments and helped with class projects (\approx 25 students).

E0 261: Database Management Systems, IISc

Oct 2018

Covered papers on Google's Spanner and Apache Giraph. (\approx 40 students).

[CV compiled on 2025-12-09]