Gauri Joshi

http://andrew.cmu.edu/user/gaurij
Google Scholar

Current Research Interests

Routing and Load balancing algorithms for LLM Inference Optimization, Distributed training and fine-tuning of large models, Reinforcement learning and Multi-armed bandit algorithms

Experience and Education

Carnegie Mellon University Pittsburgh PA
Associate Professor (without tenure) of Electrical and Computer Engineering — Jul 2022 – present
Assistant Professor of Electrical and Computer Engineering — Sept 2017 – Jul 2022

Microsoft M365
Visiting Researcher — Jan 2022 – Mar 2023

IBM T. J. Watson Research Center
Research Staff Member — July 2016 – Aug 2017

Massachusetts Institute of Technology — Sept 2012– Jun 2016
Ph.D and M.S in Electrical Engineering & Computer Science — **Best MS Thesis Award**

Indian Institute of Technology Bombay — Jul 2005–Jun 2010
B. Tech & M. Tech in Electrical Engineering; GPA 9.77/10.0 — **Institute Gold Medal**

Selected Awards and Honors

- CMU Philip and Marsha Dowd faculty fellowship for exceptional educational contributions 2025
- **IEEE Goldsmith Lecturer Award** from the Information Theory Society 2025
- Google Research Scholar Award 2023
- Named as one of **MIT Technology Review's 35 innovators under 35** 2022
- Office of Naval Research Young Investigator Award 2022
- CMU College of Engineering Dean's Early Career Fellowship 2022
- **Best Paper Award** at ACM MobiHoc 2022
- **NSF CAREER Award** 2021
- **Best Paper Award** at ACM SIGMETRICS 2020
- **Distinguished Student Paper Award**, NeurIPS FedML Workshop 2019
- **Qualcomm Innovation Fellowship**, awarded to my students Jianyu and Ankur 2019
- IBM Faculty Research Award 2018
- Rising Stars in EECS workshop invited participant 2015
- William Martin Memorial Award for **Best Masters Thesis** in Computer Science, MIT 2012
- Morris Joseph Levin Award for **Outstanding Thesis Presentation**, MIT 2012
- Claude E. Shannon Research Assistantship, MIT 2015–2016

- Schlumberger **Faculty for the Future Fellowship** 2011–2015
- Irwin and Joan Jacobs Presidential Fellowship, MIT 2010–2011
- **Institute Gold Medal** for highest GPA in the undergraduate class, IIT Bombay 2010
- Institute Silver Medal for highest GPA in the masters class, IIT Bombay 2010
- Best student in Communications & Signal processing Award, IIT Bombay 2009
- Selected among top 50 students in India for the International Chemistry Olympiad camp 2005

Selected Publications

Summary: Published more than >100 research papers, including >20 journal papers and >50 peer-reviewed conference papers, and 2 US patents. According to Google Scholar, my work has been cited $>18,000$ times with an h-index of 44.

Papers with more than 100 **citations** (full list here):

[21] P. Kairouz, B. McMahan, J. Wang, G. Joshi, et al "Advanced and Open Problems in Federated Learning", *Foundations and Trends in Machine Learning*, June 2021, arXiv:1912.04977, 8752 citations

[20] J. Wang, Q. Liu, G. Joshi and V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization", *Proceedings of Neural Information Processing Systems (NeurIPS)*, Dec 2020, arXiv:2007.07481, 1916 citations

[19] Y. Cho, J. Wang and G. Joshi "Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies", *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Apr 2022, arXiv:2010.01243, 896 citations

[18] J. Wang and G. Joshi, "Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms", *Journal on Machine Learning Research (JMLR)*, Sept 2021, arXiv:1808.07576, 700 citations

[17] J. Wang, Z. Charles, Z. Xu, G. Joshi, HB. McMahan, M. Al-Shedivat, et al, "A Field Guide to Federated Optimization", July 2021, arXiv:2107.06917, 478 citations

[16] J. Wang, G. Joshi, "Adaptive Communication Strategies to Achieve the Best Error-Runtime Trade-off in Local-Update SGD", *SysML Conference*, Mar 2019, arXiv:1810.08313, 283 citations

[15] G. Joshi, Y. Liu, E. Soljanin, "On the Delay-Storage Trade-off in Coded Distributed Storage Systems", *IEEE Journal on Selected Areas of Communications*, volume 32, number 5, May 2014, arXiv:1305.3945, 228 citations

[14] S. Dutta, G. Joshi, S. Ghosh, P. Dube, P. Nagpurkar, "Slow and Stale Gradients Can Win the Race: Error-Runtime Trade-offs in Distributed SGD", *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Apr 2018, 222 citations

[13] J. Wang, A. Sahu, G. Joshi, and S. Kar, "MATCHA: Speeding up Decentralized SGD via Matching Decomposition Sampling", *Proceedings of the Neural Information Processing Systems (NeurIPS) Federated Learning workshop*, Dec 2019, **Distinguished Student Paper Award**, arXiv:1905.09435, 214 citations

[12] YJ Cho, A Manoel, G Joshi, R. Sim, D. Dimitriadis, "Heterogeneous Ensemble Knowledge Transfer for Training Large Models in Federated Learning", *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, July 2022, arXiv:2204.12703, 180 citations

[11] D. Wang, G. Joshi, G. Wornell, "Using Straggler Replication to Reduce Latency in Large-scale Parallel Computing", *SIGMETRICS Workshop on Distributed Cloud Computing*, Jun 2015, 178

citations

[10] A. Mallick, U. Sheth, G. Palanikumar, M. Chaudhari, and G. Joshi, "Rateless Codes for Near-Perfect Load Balancing in Distributed Matrix-Vector Multiplication", *Proceedings of ACM SIGMETRICS*, June 2020, arXiv:1804.10331 **Best Paper Award**, 174 citations

[9] G. Joshi, E. Soljanin, G. Wornell, "Efficient Redundancy Techniques to Reduce Latency in Cloud Systems", *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, volume 2, issue 2, May 2017, arXiv:1508.03599, 167 citations

[8] D. Jhunjhunwala, A. Gadhikar, G. Joshi, Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, arXiv:2102.04487, 157 citations

[7] A. Jiang, D. L-K Wong, G. Zhou, D. G. Andersen, J. Dean, G. R. Ganger, G. Joshi, M. Kaminksy, M. Kozuch, Z. C. Lipton, P. Pillai, "Accelerating Deep Learning by Focusing on the Biggest Losers", Oct 2019, arXiv:1910.00762, 152 citations

[6] D. Wang, G. Joshi, G. Wornell, "Efficient Straggler Replication in Parallel Computing ", *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 2019 arXiv:1503.03128, 151 citations

[5] J. Wang, Q. Liu, G. Joshi and V. Poor, "A Novel Framework for the Analysis and Design of Heterogeneous Federated Learning", *IEEE Transactions on Signal Processing*, Aug 2021, https://ieeexplore.ieee.org/abstract/document/9521822, 130 citations

[4] G. Joshi, Y. Liu, E. Soljanin, "Coding for Fast Content Download", *Allerton Conference on Communication, Control and Computing*, Oct 2012, 114 citations

[3] A Mallick, K Hsieh, B Arzani, G Joshi, "Matchmaker: Data Drift Mitigation in Machine Learning for Large-Scale Systems", *Proceedings of Machine Learning and Systems (MLSys)*, May 2024, paper link, 111 citations

[2] YJ. Cho, L. Liu, Z. Xu, A. Fahrezi, G. Joshi, "Heterogeneous LoRA for federated fine-tuning of on-device foundation models", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov 2024, https://arxiv.org/pdf/2401.06432, 106 citations

[1] S. Khodadadian, P. Sharma, G. Joshi, S. T. Maguluri, "Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling", *Proceedings of the Conference on International Conference on Machine Learning (ICML)*, **long talk**, June 2022, https://arxiv.org/abs/2206.10185, 106 citations

PhD Student and Postdoc Advising

Advised 17 PhD students (9 graduated), 1 post-doc and 25+ masters and undergraduate students.

- Ankur Mallick, now at Microsoft 2017-2022
- Jianyu Wang, now at Google 2017-2022
- Samarth Gupta (co-advised with Prof. Osman Yağan), now at Amazon 2017-2022
- Ting-Wu (Rudy) Chin (co-advised with Prof. Diana Marculescu), now at Citadel 2020-2021
- Ahmet Inci (co-advised with Prof. Diana Marculescu), now at NVIDIA 2020-2022
- Yae Jee Cho, now at Google 2019-2024
- Tuhinangshu Choudhury (co-advised with Prof. Weina Wang), now at Meta 2019-2024
- Divyansh Jhunjhunwala, now at Amazon AGI 2020-2025

- Jiin Woo (co-advised with Prof. Yuejie Chi) 2021-present
- Pranay Sharma, post-doc, faculty at CMNIDS IIT Bombay 2021-2024
- Shuli Jiang, now at Amazon AWS 2022-2025
- Neharika Jali 2022-present
- Baris Askin (co-advised with Prof. Carlee Joe-Wong) 2022-present
- Arian Raje 2024-present
- Shivam Patel 2024-present
- Anupam Nayak (co-advised with Prof. Osman Yağan) 2024-present
- He Wang (co-advised with Prof. Yuejie Chi) 2025-present
- Susana Vittoria Marques (co-advised with Dr. Fabio Coelho), CMU Portugal 2025-present

Teaching and Textbook Development

Courses Taught at Carnegie Mellon:

- *18-461/661: Introductory Machine Learning for Engineers* Spring 2019-present
  Taken by $\sim$ 100 students every semester and offered at Pittsburgh, SV and Rwanda campuses
- *18-667: Algorithms for Distributed Machine Learning* Fall 2017-present
  Advanced graduate course covering the state-of-the-art algorithms for distributed machine learning and optimization. First ever course featuring federated learning, to my knowledge.

Textbook Development: Published a book based on 18-667 lecture notes, in the Springer Nature Synthesis lecture series on Learning, Networks and Algorithms

Selected Grants and Contracts awarded to date

Overview: The total share of funding (excluding start-up funding) awarded to my group to-date exceeds $5 million, including 8 NSF Awards, ONR YIP, and industry awards from Google, IBM, Facebook, and Bosch. Selected awards and contracts are listed below.

- *CyLab IoT Enterprise Security Grant*, "Privacy-Preserving Federated LLM Training for Critical Infrastructure Applications": $50,000, Award period: Oct 2024 - Sept 2025. PI: Gauri Joshi
- *CyLab Seed Grant*, "Synergies and Trade-offs in Distributed Optimization": $50,000, Award period: April 2024 - Mar 2025. PI: Gauri Joshi, co-PI Steven Wu
- *Pennsylvania Infrastructure Technology Alliance (PITA) grant*, " Efficient AI Workload Scheduling in Datacenters using Reinforcement Learning": $89,181, Award period: Sept 2025 - Aug 2026. PI: Guannan Qu, co-PI: Gauri Joshi
- *National Science Foundation (NSF)*, "CIF: Small: Tackling Demand and Service Uncertainty in Multi-Access Parallel Computing": $564,958, Award period: Oct 2024-Sept 2027. PI: Weina Wang, co-PI: Gauri Joshi
- *National Science Foundation (NSF) CPS Frontiers*, "Software-Defined Nanosatellite Constellations: The Foundation of Future Space-Based Cyber-physical Systems": $7,000,000, Award period: June 2022-Jun 2027, PI: Brandon Lucia, co-PIs: Vyas Sekar, Swarun Kumar, Gauri Joshi, Zachary Manchester

- *Office of Naval Research (ONR) Young Investigator Award*, "Data-Aware and System-Aware Algorithms for Distributed Machine Learning": $750,000, Award period: Jan 2023 - Dec 2025. PI: Gauri Joshi
- *National Science Foundation (NSF)*, "AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE)": $1,100,000, Award period: Oct 2021- Sept 2026, led by PI Ness Shroff from Ohio State, CMU PI: Gauri Joshi
- *National Science Foundation (NSF)*, "CAREER: Frontiers of Distributed Machine Learning with Communication, Computation and Data Constraints": $650,000, Award period: Mar 2021-Feb 2026. PI: Gauri Joshi
- *National Science Foundation (NSF)*, "SHF: Medium: Collaborative Research: HERMES: On-Device Distributed Machine Learning via Model-Hardware Co-Design": $636,000, Award period: Oct 2021- Oct 2023. CMU PI: Gauri Joshi, co-PI: Virginia Smith, UT Austin PI: Diana Marculescu, co-PI: Radu Marculescu
- *Google Faculty Research Award*, "Tackling Computational Limitations in Federated Learning": $130,000 (gift funding), Award period: Jun 2021-Jun 2023. PI: Gauri Joshi
- *Facebook Faculty Research Award*, "System-aware Distributed Optimization Algorithms for Virtual Reality Applications": $25,000 (gift funding), Award period: Jun 2021-Jun 2022. PI: Gauri Joshi
- *Carnegie Bosch Institute Research Award*, "Scheduling and Queueing Algorithms for Resource-sharing in Federated Learning": $125,000, Award period: July 2021-July 2022. PI: Gauri Joshi, co-PI: Weina Wang
- *National Science Foundation (NSF)*, "CIF: Small: Efficient Sequential Decision-Making and Inference in the Small Data Regime": $500,000, Award period: Oct 2020-Sept 2023. PI: Gauri Joshi, co-PI: Osman Yagan
- *Lawrence Livermore National Lab DoE sub-contract*, "Explainable and Small Data Machine Learning for Accelerating Feedstock Optimization": $359,462, Award period: Oct 2018-Sept 2021. PI: Gauri Joshi
- *National Science Foundation (NSF)*, "CRII: CIF: Unifying Scheduling and Algorithmic Techniques to Speed Up Distributed Stochastic Gradient Descent": $175,000, Award period: Mar 2019-Oct 2020. PI: Gauri Joshi
- *Carnegie Bosch Institute Grant*, "Privacy-Preserving Inference and Decision-Making with IoT Data": $250,000, Award period: Jan 2019-Jun 2021, PI: Osman Yağan, co-PI: Gauri Joshi
- *Bosch Research RTC Pittsburgh*, "Privacy-Preserving Inference and Decision-Making with IoT Data": $20,000 (gift funding), Dec 2018, PI: Osman Yağan, co-PI: Gauri Joshi
- *IBM Faculty Research Award*, "Fast Distributed Machine Learning with Slow and Stale Updates": $40,000 (gift funding), Award period: Jun-Dec 2018. PI: Gauri Joshi
- *National Science Foundation (NSF)*, "CIF: EAGER: Statistical Inference and Decision-Making with Sequential Samples" $100,527, Award period: Aug 2018- July 2019, PI: Osman Yağan, co-PI: Gauri Joshi

Selected Invited and Keynote Talks

Keynote Talk, IEEE CAMAD Conference, Tempe — Oct 2025

IEEE Goldsmith Award Lecture, North American School of Information Theory — June 2025

Keynote Talk, Federated Learning Workshop at KDD 2025, Toronto	July 2025
Keynote Talk, FedVision Workshop at CVPR	June 2025
Invited Talk, Information Theory and Applications Workshop, San Diego	Feb 2025
Keynote Talk, FL in the Age of Foundation Models workshop, Vancouver	Dec 2024
Keynote Talk, International Symposium on Distributed Computing (DISC)	Oct 2024
Tutorial Talk, North American School of Information Theory, Ottawa Canada	July 2024
Conference on Information Science and Systems (CISS)	Mar 2024
Plenary Talk, Texas Colloquium on Distributed Learning, Rice University	Sept 2023
TTIC Chicago Workshop on New Frontiers in Federated Learning	Sept 2023
Keynote Talk, FedVision, Federated Learning Workshop at CVPR	June 2023
Keynote Talk at Google's Federated Learning Workshop	Oct 2022
Invited talk, C3AI DTI seminar	Dec 2022
Invited talk, FLOW seminar	Nov 2022
University of Minnesota ML seminar	May 2022
NSF AI-Edge Institute Seminar	May 2022
University of Massachusetts Amherst, Machine Learning and Friends Lunch	Apr 2022
Texas A&M Computer Engineering and Systems Seminar	Feb 2022
Federated Learning One World (FLOW) Seminar	Oct 2021
Stochastics Networks, Applied Probability and Performance (SNAPP) Seminar	June 2021
Keynote Talk, SIGMETRICS Distributed Cloud Computing workshop	June 2021
Academia Seminar, Microsoft Research Redmond	May 2021
Communication and Signal Processing Seminar, University of Michigan	March 2021

Selected Professional Service and Leadership

Tutorials co-chair, ACM SIGMETRICS	2026
Program Co-Chair, ACM MobiHoc	2026
Program Co-Chair, MLSys	2025
Area Chair, Neural Information Processing Systems (NeurIPS)	2023-present
Workshop co-chair, ACM MobiHoc	2024,25
Area Chair, International Conference on Machine Learning	2025-present
Co-organizer SNAPP Seminar Series	2023-24
TPC Member, International Symposium on Information Theory (ISIT)	2024
Co-organizer of the Federated Learning workshops at NeurIPS and ICML	2020–2025
Associate Editor, IEEE/ACM Transactions on Networking	2021–2025
Student Activities Chair, ACM SIGMETRICS	2022
Co-organizer WiOpT workshop on queueing theory meets reinforcement learning	2021
Co-organizer NSF TRIPODS workshop on communication-efficient distributed optimization	2021
Publicity Chair, ACM SIGMETRICS, ACM MobiHoc, MLSys Conference	
Consulting Associate Editor of the IEEE Open Journal on Signal Processing	2020–2025
Area Chair at NeurIPS, ICML and ICLR	2020-present
TPC Member/Reviewer for ACM SIGMETRICS, MobiHoc, ISIT, MLSys, ITW	2017-present