Rui Pan

ruipan.xyz/ | github.com/ruipeterpan | ruipan@princeton.edu | Google Scholar

Education

Princeton University — Princeton, NJ, USA
Ph.D. in Computer Science — *Sep 2022 – May 2027 (Expected)*
- M.A. in Computer Science, May 2025 (conferred en route to Ph.D.)
- Advisor: Prof. Ravi Netravali

University of Wisconsin-Madison — Madison, WI, USA
B.S. in Computer Science and Mathematics — *Sep 2018 – Dec 2021*
- GPA: 3.96/4.00
- Advisor: Prof. Shivaram Venkataraman

Research Interests

I am broadly interested in the intersection between systems, networks, and machine learning. My recent work has focused on **systems for efficient ML/LLM inference**, and has been published at top systems and ML venues including SOSP, NSDI, MLSys, NeurIPS, and HotNets.

Publications (*equal contributions)

[10] SpecReason: Fast and Accurate Inference-Time Compute via Speculative Reasoning. Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, Ravi Netravali. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems* **(NeurIPS '25)**. Also in *The NeurIPS 2025 Workshop on Efficient Reasoning*.

[9] Humanity's Last Exam. Center for AI Safety, Scale AI. ArXiv preprint available (2025).

[8] METIS: Fast Quality-Aware RAG Systems with Configuration Adaptation. Siddhant Ray, **Rui Pan**, Zhuohan Gu, Kuntai Du, Shaoting Feng, Ganesh Ananthanarayanan, Ravi Netravali, Junchen Jiang. In *The 31st Symposium on Operating Systems Principles* **(SOSP '25)**.

[7] Marconi: Prefix Caching for the Era of Hybrid LLMs. Rui Pan, Zhuang Wang, Zhen Jia, Can Karakus, Luca Zancato, Tri Dao, Yida Wang, Ravi Netravali. In *The Eighth Annual Conference on Machine Learning and Systems* **(MLSys '25)**. **Outstanding Paper Award, Honorable Mention**.

[6] Mowgli: Passively-Learned Rate Control for Real-Time Video. Neil Agarwal, **Rui Pan**, Francis Y. Yan, Ravi Netravali. In *The 22nd USENIX Symposium on Networked Systems Design and Implementation* **(NSDI '25)**.

[5] Optimizing Mixture-of-Experts Inference Latency Combining Model Deployment and Communication Scheduling. Jialong Li, Shreyansh Tripathi, Lakshay Rastogi, Yiming Lei, **Rui Pan**, Yiting Xia. In *submission*, arXiv preprint available (2024).

[4] Apparate: Rethinking Early Exits to Tame Latency-Throughput Tensions in ML Serving. Yinwei Dai*, **Rui Pan***, Anand Iyer, Kai Li, Ravi Netravali. In *The 30th Symposium on Operating Systems Principles* **(SOSP '24)**.

[3] Improving DNN Inference Throughput Using Practical, Per-Input Compute Adaptation. Anand Iyer, Mingyu Guan, Yinwei Dai, **Rui Pan**, Swapnil Ghandi, Ravi Netravali. In *The 30th Symposium on Operating Systems Principles* **(SOSP '24)**.

[2] Shockwave: Fair and Efficient Cluster Scheduling for Dynamic Adaptation in Machine Learning. Pengfei Zheng, **Rui Pan**, Tarannum Khan, Shivaram Venkataraman, Aditya Akella. In *The 20th USENIX Symposium on Networked Systems Design and Implementation* **(NSDI '23)**.

[1] Efficient Flow Scheduling in Distributed Deep Learning Training with Echelon Formation. Rui Pan*, Yiming Lei*, Jialong Li, Zhiqiang Xie, Binhang Yuan, Yiting Xia. In *The 21st ACM Workshop on Hot Topics in Networks* **(HotNets '22)**.

Work Experience

Google, SystemsResearch@Google Jun 2025 – Aug 2025
Student Researcher *Sunnyvale, CA, USA*

- Mentor: Prof. Arvind Krishnamurthy
- Systems for efficient LLM inference

AWS AI, Machine Learning Systems Team May 2024 – Dec 2024
Applied Scientist Intern *Santa Clara, CA, USA*

- Mentors: Dr. Zhen Jia, Dr. Zhuang Wang, and Dr. Can Karakus
- Systems for efficient LLM inference

Max Planck Institute for Informatics, Network and Cloud Systems Group Feb 2022 – Aug 2022
Research Intern *Saarbrücken, Germany*

- Advisor: Prof. Yiting Xia
- Networked systems for efficient ML and LLM training.

Professional Activities

Reviewer:

- NeurIPS 2025 Workshop on Efficient Reasoning
- KDD 2025 Workshop on Inference Optimization for Generative AI
- IEEE Transactions on Parallel and Distributed Systems
- IEEE Transactions on Networking
- IEEE Transactions on Cloud Computing
- IEEE Transactions on Mobile Computing
- IEEE Transactions on Computers
- IEEE Transactions on Machine Learning in Communications and Networking
- IEEE Communications Letters
- Computer Networks

Artifact Evaluation Committee: MLSys '23, OSDI '23, ATC '23
Student Volunteer: CMMRS '22, N2Women@SIGCOMM '22
Teaching Assistant: COS 316 Principles of Computer System Design Fa'23, COS 598D Systems and Machine Learning Sp'24

Invited Talks

- Distributed Systems Lab, University of Pennsylvania, Dec 2024
- IEEE EDS/EPS Lunch Series: Frontier AI, UCLA, Mar 2025
- ML Systems Group, UCSD, May 2025
- KDD 2025 Workshop on Inference Optimization for Generative AI, Aug 2025

Awards

- MLSys 2025 Outstanding Paper Award, Honorable Mention
- MLSys 2025 Travel Grant ($1,000)

Last Updated: Sep 2025