Rui Pan

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

Education

Princeton University — Princeton, NJ, USA
Ph.D. in Computer Science — *Aug 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 B.S. in 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 and algorithms for efficient LLM inference**, and has been published at top systems and ML venues including MLSys, NeurIPS, SOSP, NSDI, HotNets, and IEEE ToN. My research has been recognized with an MLSys Outstanding Paper Award (Honorable Mention, 2025) and a Jane Street Graduate Research Fellowship Finalist Award (2026).

First-Author Publications (*equal contributions)

[5] **Fail Fast, Win Big: Rethinking the Drafting Strategy in Speculative Decoding via Diffusion LLMs.** **Rui Pan**, Zhuofu Chen, Hongyi Liu, Arvind Krishnamurthy, Ravi Netravali. In *submission*, arXiv preprint available (2025).

[4] **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)**. **Spotlight Paper** at *The NeurIPS 2025 Workshop on Efficient Reasoning.*

[3] **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. Being integrated into SGLang!**

[2] **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)**.

[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)**.

All Publications

[11] **A benchmark of expert-level academic questions to assess AI capabilities.** Center for AI Safety, Scale AI, HLE Contributors Consortium. In *Nature* **(Nature '26)**.

[10] **Optimizing Mixture-of-Experts Inference Time via Model Deployment and Communication Scheduling.** Jialong Li, Shreyansh Tripathi, Lakshay Rastogi, Yiming Lei, **Rui Pan**, Yiting Xia. In *The IEEE/ACM Transactions on Networking* **(ToN '26)**.

[9] **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)**.

[8] **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)**.

[7] **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)**.

[6] **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)**.

Experience

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

- Mentor: Prof. Arvind Krishnamurthy.
- Topic: Systems for efficient LLM inference.
- Developed and open-sourced FailFast, a novel speculative decoding framework that leverages Diffusion Large Language Models (dLLMs) as drafters to achieve lossless acceleration of autoregressive LLMs.
- Achieved up to 5× speedup over vanilla decoding and 2× speedup over state-of-the-art baselines like EAGLE-3 across diverse benchmarks.

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.
- Topic: Systems for efficient LLM inference.
- Researched efficient prefix caching for Hybrid LLMs (e.g., Qwen3-Next, Gemini) that combine Attention layers with subquadratic layers (e.g., Mamba, Sliding Window Attention).
- Implemented and open-sourced Marconi, a radix tree–based prefix management system; contributed to its ongoing integration into SGLang (a popular LLM inference framework with 19k+ GitHub stars).
- Improved efficiency by up to 34.4× compared to state-of-the-art prefix caching systems (e.g., vLLM and SGLang).
- Received the Outstanding Paper Award (Honorable Mention) at MLSys 2025.

Princeton University Aug 2022 – Present
Graduate Research Assistant *Princeton, NJ, USA*

- Advisor: Prof. Ravi Netravali.
- Topic: Systems for efficient ML/LLM inference.
- SpecReason: Low-latency reasoning model inference paradigm that generalizes speculative decoding by relaxing per-token equivalence into semantic equivalence, achieving complementary efficiency gains.
- METIS: Low-latency RAG system that jointly adapts RAG configurations per query and schedules them.
- Apparate and E3: Low-latency ML and LLM inference via early exiting.
- Mowgli: Offline reinforcement learning for rate control adaptation in video conferencing.
- Ongoing projects: Efficient multi-round agents, efficient multimodal model inference, efficient deep research agents.

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

- Advisor: Prof. Yiting Xia.
- Topic: Networked systems for efficient ML and LLM training.
- Researched network traffic patterns of parallelization strategies in distributed deep learning training. Proposed a network abstraction for flow scheduling in large-scale ML/LLM training clusters.

Professional Activities

Reviewer:

- Conferences: MLSys 2026, ICML 2026
- Workshops: NeurIPS 2025 Workshop on Efficient Reasoning, KDD 2025 Workshop on Inference Optimization for Generative AI
- Journals: Transactions on Machine Learning Research, 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, Future Generation Computer Systems

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 Outstanding Paper Award, Honorable Mention, 2025
- MLSys Travel Grant, 2025 ($1,000)
- NeurIPS Workshop on Efficient Reasoning, Spotlight Paper, 2025
- Jane Street Graduate Research Fellowship, Finalist Award, 2026 ($5,000)

Last Updated: Feb 2026