

Rui Pan

ruipan.xyz/ | github.com/ruipeterpan | ruipan@cs.princeton.edu | (+1) 608-960-0303

EDUCATION

Princeton University

Ph.D. in Computer Science

- Advisor: Prof. Ravi Netravali

Princeton, NJ, USA

Sep 2022 – May 2027 (*Expected*)

University of Wisconsin-Madison

B.S. in Computer Science and Mathematics

- GPA: 3.96/4.00
- Advisor: Prof. Shivaram Venkataraman

Madison, WI, USA

Sep 2018 – Dec 2021

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**.

PUBLICATIONS (*EQUAL CONTRIBUTIONS)

[7] **Marconi: Prefix Caching for the Era of Hybrid LLMs.** Rui Pan, Zhuang Wang, Zhen Jia, Can Karakus, Yida Wang, Luca Zancato, Tri Dao, Ravi Netravali. In *submission*, arXiv preprint available soon.

[6] **Tarzan: Passively-Learned Real-Time Rate Control for Video Conferencing.** Neil Agarwal, Rui Pan, Francis Y. Yan, Ravi Netravali. In *submission*, arXiv preprint available.

[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.

[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

Amazon, AWS AI Team

Applied Scientist Intern

- Mentors: Dr. Zhen Jia, Dr. Zhuang Wang, and Dr. Can Karakus

May 2024 – Dec 2024

Santa Clara, CA, USA

Max Planck Institute for Informatics

Research Intern

- Advisor: Prof. Yiting Xia

Feb 2022 – Aug 2022

Saarbrücken, Germany

PROFESSIONAL ACTIVITIES

Teaching Assistant: COS 316 Principles of Computer System Design Fa'23, COS 598D Systems and Machine Learning Sp'24

Student Volunteer: CMMRS '22, N2Women@SIGCOMM '22

Artifact Evaluation Committee: MLSys '23, OSDI '23, ATC '23