Rui Pan

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

Education

Princeton University Princeton, NJ, USA
Ph.D. in Computer Science *Sep 2022 – May 2027 (Expected)*

- 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 networked systems and machine learning. My past work was in building **systems and networks for machine learning and video applications**.

Publications (*equal contributions)

[**4**] Apparate: Rethinking Early Exits to Tame Latency-Throughput Tensions in ML Serving. Yinwei Dai*, **Rui Pan***, Anand Iyer, Kai Li, Ravi Netravali. In *submission.*

[**3**] Fast & Efficient DNN Inference Using Practical Early-Exit Networks. Anand Iyer, Swapnil Ghandi, Mingyu Guan, Yinwei Dai, **Rui Pan**, Ravi Netravali. In *submission.*

[**2**] Shockwave: Fair and Efficient Cluster Scheduling for Dynamic Adaptation in Machine Learning. Pengfei Zheng, **Rui Pan**, Tarannum Khan, Shivaram Venkataraman, Aditya Akella. In *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 *21st ACM Workshop on Hot Topics in Networks* **(HotNets '22)**.

Research Experience

Graduate Research Assistant @ Princeton University Princeton, NJ, USA
Project 1: Practical, Accuracy-Preserving Early Exiting for DNN Serving *Sep 2022 - Present*

- Advisors: Prof. Ravi Netravali, Prof. Kai Li, and Prof. Anand Iyer (Georgia Tech)
- Designed a system that automatically applies and manages early exits (EEs) in serving ML models, whereby certain inputs can exit with results at intermediate layers.
- Integrated EE into popular CV and NLP workloads; Studied, implemented, and evaluated the batching logic of common serving platforms like Clockwork, TF-Serve, and Triton Inference Server; Developed and open-sourced a profiler for profiling the architecture and training/inference speed of DNNs.
- Across diverse workloads, Apparate lowers median response latencies by up to 91.5% without affecting throughput or violating tight accuracy constraints.
- Submitted a conference paper as the first author.

Project 2: Towards Practical Data-Driven Bitrate Adaptation in Video Conferencing *Dec 2022 - Present*

- Advisors: Prof. Ravi Netravali, Dr. Francis Yan (Microsoft)
- Working on making bitrate selection in videoconferencing less conservative and quicker to adapt to bandwidth changes via offline reinforcement learning, which learns behaviors using only logged data without further environment interaction.
- Implemented online RL-based bandwidth estimation schemes to serve as a baseline and to motivate the need for practical reinforcement learning in networked applications.
- Preparing the submission of a conference paper as the second author.

Max Planck Institute for Informatics — Feb 2022 – Aug 2022

Research Intern — *Saarbrücken, Germany*

- Advisor: Prof. Yiting Xia, Dr. Jialong Li
- Researched network traffic patterns of emerging parallelization paradigms in distributed deep learning training. Proposed a network abstraction for flow scheduling in ML clusters.
- Published a workshop paper as the first author at HotNets '22.

Undergraduate Research Assistant @ UW-Madison — Madison, WI, USA

Project 1: Fair and Efficient Resource Allocation for DNN Training in GPU Clusters — *Mar 2021 – Dec 2021*

- Advisors: Prof. Shivaram Venkataraman, Prof. Aditya Akella (UT Austin), and Dr. Pengfei Zheng
- Developed a policy to co-optimize long-term fairness and efficiency of the scheduling/resource allocation of resource-adaptive deep learning training workloads in large-scale multi-tenant GPU clusters.
- Implemented and integrated the novel allocation policy into Gavel [OSDI '20], an existing scheduling framework. Implemented the mechanism to support dynamic adaptation (e.g., batch size scaling) of training workloads in Gavel.
- Implemented dynamic optimizations, e.g., Accordion [MLSys '21] & Gradient Noise Scale [arXiv '18], for common DNN training workloads to increase the training efficiency without loss of accuracy.
- Achieved 1.3x efficiency win and 2x fairness win at the same time over state-of-the-art scheduling policies (Themis [NSDI '19], Gavel [OSDI '20], AlloX [EuroSys '20]) on a trace of real-world workloads in both large-scale simulations and physical experiments.
- Published a conference paper as the second author at NSDI '23.

Project 2: How Structured Backpropagation Pruning Improves Deep Learning Clusters — *Jun 2020 – Feb 2021*

- Advisor: Prof. Shivaram Venkataraman
- In this work, we systematically control the amount of backpropagation at individual workers in distributed DNN training. This technique, Structured Backpropagation Pruning (SBP), simultaneously reduces network bandwidth, compute utilization, and memory use while preserving model quality.
- Developed an iteration-level cluster scheduler by extending existing frameworks such as PyTorch Elastic and BytePS [OSDI '20] to capitalize on the resources saved by SBP. The scheduler supports fine-grained iteration-level scheduling, different communication protocols, frequent checkpointing, and worker migration with low overhead.
- Used Microsoft Azure to develop, deploy, and modify existing code bases. Profiled common workloads to identify the communication bottlenecks in distributed DNN training and filed issue reports to open-source frameworks.

Wisconsin Institute for Discovery — Jan 2020 – Mar 2021

Undergraduate Research Assistant — *Madison, WI, USA*

- Advisors: Dr. Steven Wangen and Prof. Michael Ferris
- Proposed Dairy Brain, an analytics platform for evaluating and predicting the performance of dairy cows by aggregating large quantities of dairy data.
- Developed, deployed and maintained a data warehouse, Agricultural Data Hub (AgDH), for the collection, storage, homogenization, entity matching, and distribution of dairy farm's feeding, milking, and management data in a series of PostgreSQL data marts. Assisted with the implementation of the data pipeline using Apache Airflow.
- Presented our poster at the 3rd Wisconsin Institute for Discovery (WID) Research Symposium and in outreach meetings for the local dairy industry.

Professional Activities

Teaching Assistant: COS 316 Fa'23
Student Volunteer: CMMRS '22, N2Women@SIGCOMM '22
Artifact Evaluation Committee: MLSys '23, OSDI '23, ATC '23

Technical Skills

Languages: Python, Java/C#, C/C++, SQL, JavaScript, HTML/CSS, R
Frameworks and Tools: PyTorch, CUDA, Docker, PostgreSQL, OpenMP, MPI, Apache Spark, Microsoft Azure