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 building **systems and networks** for **machine learning and video applications**.

Publications (*equal contributions)

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

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

[**3**] 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)**.

[**4**] 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

Princeton University — Sep 2022 - Present
Graduate Research Assistant — *Princeton, NJ, USA*
- Advisor: Prof. Ravi Netravali

Max Planck Institute for Informatics — Feb 2022 – Aug 2022
Research Intern — *Saarbrücken, Germany*
- Advisor: Prof. Yiting Xia
- Researched network flow scheduling strategies for emerging parallelization paradigms in distributed deep learning training.

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
- 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.

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

Relevant Courses

Graduate CS: Computer Vision, Distributed Systems, Networked Systems, Programming Languages
Undergraduate CS: Advanced Operating Systems, Algorithms, Artificial Intelligence, Big Data Systems (audited), Bioinformatics, Computer Architecture, Computer Vision, Database Systems, Data Analysis, High Performance Computing, Networks, Operating Systems, Theory of Computing
Undergraduate Math: Calculus, Combinatorics, Discrete Math, Math in Data Science, Numerical Linear Algebra, Probability

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