Peiran Qin

Pre-Doctoral M.S. Student
University of Chicago
5035 East End, Chicago, IL 60615

Email: peiranqin2001@gmail.com
Phone: +1 (773) 812 5328
https://qinpr.github.io/

I'm activately looking for **2025 Fall Ph.D. Opportunities**!

Research Interests

Areas: Operating Systems, Storage Systems, Distributed Systems and their intersections with Machine Learning.

Focus: Building more performant systems in the context of:

1) High Resouce Efficiency (e.g. prefetch and cache eviction co-design for higher cache efficiency)
2) Low Latency (e.g. mitigating tail latencies in cloud storage systems)
3) Workload Awared (e.g. workload characterization and prediction for distributed ML clusters)

Education

University of Chicago

Pre-Doctoral M.S. in Computer Science; GPA: 3.81/4.00 — Sep, 2023 - Expected Dec, 2024

- Advisor: Haryadi S. Gunawi
- Merit-based Scholarship recipient (4 / CS class of 2024)

Chinese University of Hong Kong, Shenzhen

B.Eng. in Computer Engineering; GPA: 3.53/4.00 — Sep, 2019 - Jul, 2023

- Advisor: Yeh-Ching Chung
- First Class Academic Honours recipient

Publications

In Submission '24 — Anonymous Author(s). **Heimdall: Accurate and Efficient I/O Admission Policy with Extensive Machine Learning Pipeline.**
Manuscript is available upon request

In Submission '24 — Anonymous Author(s). **Storage Prefetching Needs Feedback: An Efficient Prefetching/Eviction Co-Design with Small Cache Footprints.**
Manuscript is available upon request

Research Experience

University of Chicago; Advisor: Haryadi S. Gunawi

Research Assistant, UCARE Group — Apr, 2023 - Present

Cache Prefetching & Eviction Co-Design

- Developed a co-designed policy for cache prefetching and eviction, achieving a 62% reduction in backend load compared to the state-of-the-art.
- Implemented the system on multiple platforms including CacheLib, Linux kernel, Ceph.
- Implemented and characterized SOTA prefetchers (e.g. Pythia) and eviction policies (e.g. SIEVE, S3FIFO).
- Majorly contribute to structure, organize, and conduct the evaluation experiments.

(Paper in submission '24)

I/O Admission Policy

- Built an efficient I/O admission policy by innovatively combining storage domain knowledge with the extensive ML pipeline. Our system surpasses the SOTA by 15%-35% in terms of average I/O latency.
- Implemented the joint-inference which can maintain the scalability under 8x heavier workload.
- Hacked the deployment of our system on Linux kernel level and C client level.
- Implemented a SSD wear-leveling emulator which is cheap, scalable and extensible, based on FEMU.
- Created artifacts on Chameleon for public reproduction.

(Paper in submission '24)

Microsoft Research, Asia; Mentor: Lei Qu

Research Intern, Networking Research Group May, 2022 - Dec, 2022

Workload Characterization & Prediction for ML Clusters Design

- One of the major contributors to the AISim project, supporting ML distributed systems design by workloads analysis.
- Completed the design, implementation, and validation to extract the general workload characteristic from first-part clusters and proposed a pipeline to predict the workload's distribution.
- Created a workload prediction model with high goodness-of-fit (p-value=0.3) and generalizability.

(Research report delivered to Microsoft Engineering Group '22)

Workload Profiling in ML Clusters

- One of the major contributors to DMLProfiling project, focusing on ML distributed systems profiling.
- Characterized the resource patterns of various large ML models running in Microsoft Azure clusters.
- Devised a high-precision solution for inferring the job types and model types (ResNet, BERT, ViT, etc) of jobs run on clusters by analyzing the key resources usage, with the accuracy reaching over 99%.

(Research report delivered to Microsoft Engineering Group '22)

Alibaba Group; Advisor: Yeh-Ching Chung

Student Research Assistant Dec, 2021 - May, 2022

Graph Computing Sytems Optimization

- Worked on the large-scale sub-graph matching systems, focusing on reducing memory footprint.
- Optimized the design by replacing the stale tabular-based catalog with graph-based representation equipped with dynamic programming, which reduces the memory usage and improves the efficiency.
- Proposed a solution for encoding and decoding sub-graphs with reduced memory footprint.

(Paper accepted at ATC '23)

Other Experience

Repeto Program, Summer of Reproducibility Fellowship

- *Funded by NSF*; Contributor Jun, 2024 - Aug, 2024

Zhongshen Agricultural Innovation Technology Co., Ltd

- Software Development Engineer May, 2021 - Dec, 2021

Technical Skills

Linux Kernel:	Experience of block layer, vfs, and syscalls hacking in Linux-6.0
Programming Language:	C, C++, Python, Rust, Java, C#, Verilog HDL, SQL
Distributed Systems:	RocksDB, Ceph
ML Framework:	PyTorch, TensorFlow
Containerization:	Docker
Parallel Computing Libraries:	CUDA, MPI, Pthread, OpenMP, NCCL
Other	LaTeX

References

Haryadi S. Gunawi
Associate Professor
University of Chicago
haryadi@cs.uchicago.edu

Yongqiang Xiong
Principal Researcher
Microsoft Research, Asia
yqx@microsoft.com

Yeh-Ching Chung
Professor
Chinese University of Hong Kong, Shenzhen
ychung@cuhk.edu.cn