Peiran Qin

Software Systems Engineer
MangoBoost, Inc.
14240 SE 6th St, Bellevue, WA 98007

Email: peiranqin2001@gmail.com
Phone: +1 (773) 812 5328

Areas of Expertise

Areas: Systems for LLM Inference and Training. Operating Systems, Storage Systems, and Distributed Systems.

Focus: Building more performant ML systems in the context of:

1) High Resouce Efficiency (e.g. optimizing cache eviction policy 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: 3.86/4.00 — Sep, 2023 - Dec, 2024

- Advisor: Haryadi S. Gunawi
- Recipient of a $38,000 Merit-based Scholarship (4 students, CS class of 2024)

Chinese University of Hong Kong, Shenzhen

B.Eng. in Computer Engineering; First-Class Honours — Sep, 2019 - Jul, 2023

- Advisor: Yeh-Ching Chung

Publications

EuroSys'25 [Acceptance Rate: 8.2%]	Daniar H. Kurniawan, Rani Ayu Putri, Peiran Qin, Kahfi S. Zulkifli, Ray AO Sinurat, Janki Bhimani, Sandeep Madireddy, Achmad Imam Kistijantoro, and Haryadi S. Gunawi. **Heimdall: Accurate and Efficient I/O Admission Policy with Extensive Machine Learning Pipeline.**
In Submission '25	Anonymous Author(s). **Storage Prefetching Needs Feedback: An Efficient Prefetching/Eviction Co-Design with Small Cache Footprints.** *Manuscript is available upon request*

Work Experience

MangoBoost, Inc. (Series-A Startup) — Bellevue, WA

Software Enigneer, LLM Inference Systems — Jan, 2025 - Present

Systems for LLM Inference [Product Link]

- Improved distributed inference on 64x MI355X GPUs in partnership with AMD, achieving 93K tokens/s on Llama2-70B FP4, performance comparable to B200 GPUs.
- Optimized network stack for the first heterogeneous cluster MLPerf benchmarking submission, enabling 25% higher throughput and linear scaling.
- Customized and optimized vLLM core scheduler, reducing latency by 2x on Qwen-2.5-VL-72B.
- Characterized KV cache DRAM and storage offloading strategies in LMCache, delivered comprehensive overhead analysis for future improvement.
- Developed key features of LLMBoost Inference Engine (core-engine, networking, and auto-config tuning algorithms). Delivered up to 7.1x higher throughput on Llama4-Scout.

(Media Mentions: [AMD], [Forbes], [IEEE], [SuperMicro])

Systems for LLM Training

- Optimized on the NCCL and Megatron optimization. Offered near-linear throughtput scaling from 8x MI300X GPUs to 32x MI300X GPUs on Llama2-70B-LoRA Finetuning.
- Optimized on the context parallelism computation-communication overlapping, achieved 11% lower lantency than previous optimal parallelism strategy.
- Delivered the first-ever multi-node (16x & 32x) AMD GPUs result in MLPerf, competitive to H200 cluster.

(Media Mentions: [AMD], [Dell], [Forbes], [ZDNET])

University of Chicago; Advisor: Haryadi S. Gunawi — Chicago, IL
Research Assistant, UCARE Group — Apr, 2023 - Jan, 2025

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 6.0.0 kernel, Ceph.
- Implemented and characterized SOTA prefetchers (e.g. Pythia) and eviction policies (e.g. SIEVE, S3FIFO).

(Paper in submission '25)

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.

(Paper accepted at EuroSys'25)

Microsoft Research, Asia; Mentor: Lei Qu — Beijing, China
Intern, Networking Research Group — May, 2022 - Dec, 2022

Workload Characterization & Prediction for ML Clusters Design

- One of the major contributors to the AISim project, supported MLsys 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 — Hangzhou, China
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)

Technical Skills

Systems for LLM:	vLLM, SGLang, Megatron, Mango LLMBoost
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

References

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

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

Juncheng Yang
Assistant Professor
Harvard University
juncheng@g.harvard.edu