

Lequn Chen 陈乐群

✉ lqchen@cs.washington.edu • 🌐 abcdabcd987.com • 📍 Seattle, WA

Education

University of Washington

Ph.D., Computer Science, Advisor: **Arvind Krishnamurthy**

Seattle, WA
Sep 2018–(Expected) Mar 2024

Thesis: Multi-tenant Machine Learning Model Serving Systems on GPU Clusters

Research Domain: Machine Learning Systems, Distributed Systems, Operating Systems

ACM Honors Class, Shanghai Jiao Tong University

Bachelor, Computer Science, Advisors: **Weinan Zhang**, **Gui-Rong Xue**, and **Yong Yu**

Shanghai, China

Sep 2014–Jun 2018

Cornell University

Visiting Research Intern, Advisors: **Emin Gün Sirer** and **Kai Mast**

Ithaca, NY

Jul 2017–Dec 2017

Research and Selected Publications

My Ph.D. research focuses on improving model serving efficiency in a multi-tenant setting. The following work explores related research problems, including: (1) GPU batching efficiency under latency constraints, (2) handling bursty requests, (3) GPU consolidation and autoscaling, (4) cost amortization across tenants, and (5) scalability.

Punica: Serving Multiple LoRA Finetuned LLMs at the Cost of One

(in review, arXiv 2310.18547)

Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, Arvind Krishnamurthy

🔗punica-ai/punica

○ HackerNews front page; 2023 Madrona Prize; a16z; adopted by several companies; 400+ stars in two weeks.

○ Achieved **12x throughput** compared to state-of-the-art LLM serving systems without latency sacrifice.

○ Created a new paradigm of serving any number of finetuned large language models at the cost of one.

Symphony: Optimized Model Serving using Centralized Orchestration

(in review, arXiv 2308.07470)

Lequn Chen, Weixin Deng, Anirudh Canumalla, Yu Xin, Matthai Philipose, Arvind Krishnamurthy

🔗abcdabcd987/nexus1b

○ Achieved **6x goodput** compared to state-of-the-art DNN model serving systems. Increased batch size by **3x**.

○ Reduced required number of GPUs by **60%** by resource pooling across tenants. Eliminated over-provisioning.

○ Crafted a centralized scheduler that scales to **15 million RPS** on a single server.

○ Designed a practical and elegant fault-tolerance protocol that delivers millisecond-scale recovery.

Nexus: A GPU Cluster Engine for Accelerating DNN-Based Video Analysis

(SOSP'19)

Haichen Shen, Lequn Chen, Yuchen Jin, Liangyu Zhao, Bingyu Kong, Matthai Philipose, Arvind Krishnamurthy, Ravi Sundaram

🔗uwsaml/nexus

○ Prior state-of-the-art for multi-tenant DNN model serving systems. 80+ citations.

Atom: Low-bit Quantization for Efficient and Accurate LLM Serving

(in review, arXiv 2310.19102)

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, Baris Kasikci

○ Improved text generation throughput by **7.7x** with W4A4 compared to FP16, 5.5x to W4A16, and 2.5x to W8A8.

○ Reduced perplexity by half compared to state-of-the-art quantization schemes, adding only 10% compared to FP16.

Industry Experience

Google

Software Engineering Intern, Vertex AI team, Disaggregated Model Serving project.

Seattle, WA

Jun 2022–Sep 2022

○ Implemented a proof-of-concept of disaggregated model serving, projecting a 3x reduction in total cost of ownership.

Microsoft Research

Research Intern, WatchFor project

Redmond, WA

Jun 2021–Sep 2021

○ Prototyped a transfer learning tool that finds the accuracy-latency Pareto frontier 8x faster using Once-For-All model.

Google

Software Engineering Intern, Cache Invalidation and Notification team

Kirkland, WA

Jun 2019–Sep 2019

○ Added a new feature to the Memcache service: replication. Implemented with 10k lines of C++ code.

Technical Skills

System Programming: C++, Python, Rust, asynchronous, multithread, multiprocessing, distributed, RDMA

Machine Learning: PyTorch, Numpy, Matplotlib, TensorFlow, JAX, XLA, HuggingFace

Full Stack: Web Frontend, Backend, PostgreSQL, Grafana, Docker, Kubernetes, CI/CD, Sysadmin, Security