Lequn Chen 陈乐群

chenlequn22@gmail.com • abcdabcd987.com • Seattle, WA

Education

University of Washington **Seattle, WA**
Ph.D. and Master, Computer Science, Advisor: **Arvind Krishnamurthy** *Sep 2018–Mar 2024*
Research Domain: Machine Learning Systems, Distributed Systems, Operating Systems

Shanghai Jiao Tong University (ACM Honors Class) **Shanghai, China**
Bachelor, Computer Science, Advisors: **Weinan Zhang**, **Gui-Rong Xue**, and **Yong Yu** *Sep 2014–Jun 2018*

Cornell University **Ithaca, NY**
Visiting Research Intern, Advisors: **Emin Gün Sirer** and **Kai Mast** *Jul 2017–Dec 2017*

Research and Selected Publications

My Ph.D. research focuses on model serving efficiency in a multi-tenant setting, tackling the following research problems: (1) GPU batching efficiency under latency constraints, (2) Request scheduling, (3) GPU provisioning and autoscaling.

Punica: Serving Multiple LoRA Finetuned LLMs at the Cost of One (MLSys'24)
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 DNN Model Serving using Deferred Batch Scheduling (review, 2308.07470)
Lequn Chen, *Weixin Deng, Anirudh Canumalla, Yu Xin, Danyang Zhuo, Matthai Philipose, Arvind Krishnamurthy* abcdabcd987/nexuslb
- 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* uwsampl/nexus
- Prior state-of-the-art for multi-tenant DNN model serving systems. 100+ citations.

Atom: Low-bit Quantization for Efficient and Accurate LLM Serving (MLSys'24)
Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, ***Lequn Chen****, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, Baris Kasikci* efeslab/Atom
- 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.

Work Experience

Perplexity AI **Remote**
Research Engineer *Apr 2024–Present*

Google **Seattle, WA**
Software Engineering Intern, Vertex AI team, Disaggregated Model Serving project. *Jun 2022–Sep 2022*
- Implemented a proof-of-concept of disaggreated model serving, projecting a 3x reduction in total cost of ownership.

Microsoft Research **Redmond, WA**
Research Intern, WatchFor project *Jun 2021–Sep 2021*
- Prototyped a transfer learning tool that finds the accuracy-latency Pareto frontier 8x faster using Once-For-All model.

Google **Kirkland, WA**
Software Engineering Intern, Cache Invalidation and Notification team *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, multiprocess, distributed, RDMA
Machine Learning: PyTorch, Numpy, Matplotlib, TensorFlow, JAX, XLA, HuggingFace
Full Stack: Web Frontend, Backend, PostgreSQL, Grafana, Docker, Kubernetes, CI/CD, Sysadmin