Lequn Chen 陈乐群

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

Education

University of Washington **Seattle, WA**
Ph.D., Computer Science, Advisor: **Arvind Krishnamurthy** *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
Teaching Assistant: CSE550 Computer Systems, CSE552 Distributed and Parallel Systems

ACM Honors Class, Shanghai Jiao Tong University **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*

Selected Publications

My Ph.D. research is dedicated to improving model serving efficiency in multi-tenant setting. The following work explores related research problems, including (1) GPU batching efficiency under latency constraints. (2) Handling bursty requests. (3) Resource consolidation and autoscaling. (4) Computation and storage amortization across tenants. (5) State affinity across iterations.

1. **Punica: Multi-Tenant LoRA Fine-Tuned LLM Serving** (arXiv 2310.18547)
   Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, Arvind Krishnamurthy
2. **Atom: Low-bit Quantization for Efficient and Accurate LLM Serving** (arXiv 2310.19102)
   Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, **Lequn Chen**, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, Baris Kasikci
3. **Symphony: Optimized Model Serving using Centralized Orchestration** (arXiv 2308.07470)
   Lequn Chen, Weixin Deng, Anirudh Canumalla, Yu Xin, Matthai Philipose, Arvind Krishnamurthy
4. **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

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, Security

Industry Experience

Google **Seattle, WA**
Software Engineering Intern, Vertex AI, Disaggregated model serving. *Jun 2022–Sep 2022*

- ○ Design and implementation a model rewriting tool for disaggregated serving.
- ○ Measurement of latency overhead and estimation of benefits in total cost of ownership.
- ○ Rule-based model optimization for disaggregated serving.

Microsoft Research **Redmond, WA**
*Research Intern, **WatchFor** Project* *Jun 2021–Sep 2021*

- ○ Investigated into the Pareto frontier of accuracy-latency trade-off of transfer learning models.
- ○ In-depth study of how to perform transfer learning and neural architecture search effectively and efficiently.
- ○ Explored advanced compiler optimization opportunities and challenges for the Pareto frontier models, e.g., GPU memory sharing across models, layer-based optimization caching.

Google **Kirkland, WA**
*Software Engineering Intern, **Tango** Team, Cache Invalidation and Notification* *Jun 2019–Sep 2019*

- ○ In-depth discussion of the *Virtual Object Set* feature of the next-generation Tango.
- ○ Added a new feature to the Memcache service: replication.
  - Reasoned about consistency guarantees of the new feature.
  - Implemented with 10k lines of C++ code.
  - Covered by unit tests and integration tests.