Lequn Chen 陈乐群

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

Education

University of Washington **Seattle, WA**
Ph.D., Computer Science *Sep 2018–(Expected) Mar 2024*
Thesis: Multi-tenant Machine Learning Model Serving Systems on GPU Cluster
Research Domain: Machine Learning Systems, Distributed Systems, Operating Systems
Teaching Assistant: CSE550 Computer Systems, CSE552 Distributed and Parallel Systems
Advisor: Arvind Krishnamurthy

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

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

Publications

1. **Punica: Multi-Tenant LoRA 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
5. **ADARES: Adaptive Resource Management for Virtual Machines** (arXiv 1812.01837)
   Ignacio Cano, **Lequn Chen**, Pedro Fonseca, Tianqi Chen, Chern Cheah, Karan Gupta, Ramesh Chandra, Arvind Krishnamurthy
6. **Scaling Databases through Trusted Hardware Proxies** (SysTEX'17)
   Kai Mast, **Lequn Chen**, Emin Gün Sirer

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*

- 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.
- In-depth discussion of the *Virtual Object Set* feature of the next-generation Tango.

Prior Research Experiences

Systems Lab **Cornell University**

*Visiting Research Intern, advised by **Emin Gün Sirer** and **Kai Mast*** *Jul 2017–Dec 2017*

- Worked on a database that provides blockchain-like guarantees of data integrity using *Trusted Execution Environments*.
  - Implemented large parts of the prototype on Intel SGX. Boosted the performance inside the SGX enclave.
  - Increased the throughput of multi-client read workload 30x and reduced the latency by 40%.
  - Implemented transaction support with optimistic concurrency control.
  - Optimized query optimizer and executor, reducing cost of join operation to almost constant in typical workloads.
  - Found and solved dozens of deadlocks and data races in the initial version of the code.
  - Designed benchmarks and conducted experiments on a distributed testbed.

APEX Data & Knowledge Management Lab **Shanghai Jiao Tong University**

*Undergraduate Researcher, advised by **Weinan Zhang*** *Mar 2017–Jun 2017*

- Worked on Computational Advertisement. Built a machine learning pipeline for an advertisement exchange startup.
  - Designed and trained a *Click-Through Rate* (CTR) estimation model.
  - Integrate the model with the startup's *Real-Time Bidding* (RTB) software stack.

Tianrang Network Technology Co.,Ltd **Shanghai**

*Research Intern, advised by **Gui-Rong Xue*** *Jun 2016–Mar 2017*

- Worked on a program Yi playing board game Go similar to Google DeepMind's *AlphaGo*. Yi runs Monte-Carlo tree search algorithm, deep neural network, and reinforcement learning algorithms.
  - Designed and Implemented a distributed system running both CPU and GPU workers on multiple machines.
  - Reduced the network latency and increased single-machine performance.
  - Refactored the code base. Trained and tuned neural networks. It could beat entry-level professional human players.

Teaching Experiences

CSE550 Computer Systems: Teaching Assistant *Autumn 2019*

CSE552 Distributed and Parallel Systems: Teaching Assistant *Autumn 2022*

Compilers *Spring 2017*

Student Instructor https://acm.sjtu.edu.cn/compiler2017

- Led the teaching assistant team. Re-designed assignments.
- Built a *Continuous Integration* (CI) system abcdabcd987/acm-compiler-judge, automatically testing students' new commits and updating the leaderboard.

Principle and Practice of Computer Algorithms *Summer 2016*

Student Instructor https://acm.sjtu.edu.cn/wiki/PPCA_2016

- Built an online judge system for algorithm exams.
- Led a group of students to implement simplified MapReduce and Google File System. Deployed and benchmarked them on all machines of the computer room.

C++ Programming *Autumn 2015*

Teaching Assistant https://acm.sjtu.edu.cn/wiki/Programming_2015

- Gave a lecture on how to begin C++ projects. Review and make comments to students' projects.
- Built an online judge system abcdabcd987/p2dv.in for the game bot project.
- Carefully designed two sets of homework to help students master the basic idea of OOP in C++.