Qianlin Liang

165 Brittany Manor Drive, Amherst, MA, 01002
qliang@cs.umass.edu • (413) 404-8659 • https://qianlin404.github.io/

EDUCATION

University of Massachusetts, Amherst, MA, USA

Ph.D in Computer Science Aug. 2018 – Present

- Research Direction: Distributed Systems, Cloud Computing, Edge Computing, Sustainable Computing, AI Systems
- Thesis: Rethinking the Systems and Paradigms for Cloud and Edge AI Workloads
- Expected Graduation Date: July, 2024

M.S. in Computer Science Aug. 2018 – Dec. 2020

- Cumulative GPA: 3.94/4.00

The Pennsylvania State University, University Park, PA, USA

B.S.(Hons.) in Computer Science Aug. 2012 – May 2016

- Minor in Mathematics
- Thesis: A Study of Price and Capacity Trade-Offs of Replicating Computation on the Public Cloud
- Cumulative GPA: 3.91/4.00

ACADEMIC EXPERIENCE

University of Massachusetts, Amherst

Research Assistant in the Laboratory for Advanced Software Systems Aug. 2018 – Present
Advisor: Prashant Shenoy

Advancing and transitioning the scientific foundations of performant and resilient intelligent computational and sensing services, tailored for the future tactical network edge and cloud

- Developed novel mechanisms to optimize performance and energy efficiency for Graphics Processing Unit (GPU) deployment around machine learning models
- Designed and implemented systems to provide timeliness , adaptability, fairness and energy efficiency for inference serving of ML models, utilizing state-of-the-art tools and technologies, such as PyTorch, Nvidia Triton, TensorRT, ONNX, and TVM
- Designed and trained adaptive DNN models to dynamically optimize performance, energy efficiency, and accuracy
- Developed performance model and resource management algorithms for edge cloud with specialized hardware accelerators to improve their utilization while meeting applications SLOs
- Designed a distributed machine learning inference system for resilience and fault tolerance in dynamically changing adversarial environments
- Published project findings in prestigious conferences and journals such as SEC, IoTDI, TAAS, and IISWC

Advancing sustainable computing with a software-defined energy virtualization layer for targeted energy and carbon management, promoting carbon efficiency at multiple geographical scales

- Developed a virtual energy system enabling application-level grid carbon monitoring and control of server power and battery usage
- Designed and implemented a sustainable cloud resource management system to decarbonize large-scale clusters for distributed ML training, using Kubernetes CRDs and customized controllers
- Developed an online tool for pre-deployment analysis of workloads' carbon footprint through discrete-event simulation
- Published project findings in prestigious conferences, including ASPLOS and SIGMETRICS

The Pennsylvania State University, University Park

Undergraduate Research Assistant in Computer Systems Lab May 2015 – May 2016
Advisor: Bhuvan Urgaonkar

Improving the cost-efficacy of public cloud

- Analyzed Amazon EC2 Spot market history price and developed statistic model to predict EC2 Spot market price
- Implemented controller to launch, terminate and run jobs on EC2 instances programmactically
- Designed and implemented algorithm for EC2 users to lessen their cost while maintaining high reliability
- Published project findings in prestigious conferences and workshop, including EuroSys and ICPE

INDUSTRY EXPERIENCE

Adobe Research

Research Scientist Intern in System Technology Lab May 2022 – Aug. 2022

Performance engineering for machine learning inference workloads on heterogeneous GPUs in the cloud

- Evaluated the performance impact of various optimization stages within the pipeline for different DNN models on GPUs
- Conducted comprehensive profiling of diverse DNN models under varied GPU resource allocations using MPS, gathering detailed performance data for analysis
- Designed and implemented an analytic system using a novel Graph Neural Network (GNN)-based approach to estimate DNN inference latency across various GPU resource allocations

Shanghai Rajax Information Technology Co., Ltd

Data Scientist Aug. 2016 – May 2018

Collaborate with the operations team and work towards shipping robust AI solutions

- Designed and trained machine learning models to accurately predict the delivery capacity, using TensorFlow and Scikit-learn
- Designed supply and demand pricing model to improve service quality during peak time
- Developed algorithms to cluster operating area and improve operating efficacy and efficiency

PUBLICATIONS

[1] Walid Hanafy, **Qianlin Liang**, Noman Bashir, David Irwin, Prashant Shenoy. CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing Carbon-Efficiency, In *Proceedings of ACM Special Interest Group on Measurement and Evaluation (SIGMETRICS), 2024.*

[2] **Qianlin Liang**, Walid Hanafy, Noman Bashir, David Irwin, Prashant Shenoy. Energy Time Fairness: Balancing Fair Allocation of Energy and Time for GPU Workloads, In *Proceedings of ACM/IEEE Symposium on Edge Computing (SEC) 2023*.

[3] **Qianlin Liang**, Walid A. Hanafy, Noman Bashir, Ahmed Ali-Eldin, David Irwin, Prashant Shenoy. Dělen: Enabling Flexible and Adaptive Model-serving for Multi-tenant Edge AI. In *Proceedings of IEEE/ACM Eighth International Conference on Internet-of-Things Design and Implementation (IoTDI), San Antonio May 2023*.

[4] Abel Souza, Noman Bashir, Jorge Murillo, Walid Hanafy, **Qianlin Liang**, David Irwin, Prashant Shenoy. Ecovisor: A Virtual Energy System for Carbon-Efficient Applications. In *Proceedings of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS), Vancouver, Canada, March 2023.*

[5] **Qianlin Liang**, Walid A. Hanafy, Ahmed Ali-Eldin, Prashant Shenoy. Model-driven Cluster Resource Management for AI Workloads in Edge Clouds. In *ACM Transactions on Adaptive andAutonomous Systems (TAAS) Jan 2023.*

[6] **Qianlin Liang**, Prashant Shenoy, David Irwin. AI on the Edge: Rethinking AI-based IoT Applications Using Specialized Edge Architectures. In *Proceedings of IEEE International Symposium on Workload Characterization, October 2020.*

[7] Cheng Wang, Bhuvan Urgaonkar, Aayush Gupta, **Qianlin Liang**, and George Kesidis. Exploiting Spot and Burstable Instances for Improving the Cost-efficacy of In-Memory Caches on the Public Cloud. In *Proceedings of the European Conference on Computer Systems (EUROSYS 2017), Belgrade, Serbia, April 2017.*

[8] Cheng Wang, **Qianlin Liang**, and Bhuvan Urgaonkar. An Empirical Analysis of Amazon EC2 Spot Instance Features Affecting Cost-effective Resource Procurement. In *ACM/SPEC International Conference on Performance Engineering (ICPE 2017), L'Aquila, Italy, April 2017.*

[9] **Qianlin Liang**, Cheng Wang, and Bhuvan Urgaonkar. Spot Characterization: What are the Right Features to Model? In *Proceedings of the First International Workshop on System Analytics and Characterization (SAC 2016), co-located with ACM SIGMETRICS 2016, Antibes Juan-les-pines, France, June 2016.*

CONFERENCE PRESENTATIONS

- *Energy Time Fairness: Balancing Fair Allocation of Energy and Time for GPU Workloads*. Oral presentation delivered at ACM/IEEE Symposium on Edge Computing (SEC) Wilmington, DE, Dec 06, 2023.

- *Dělen: Enabling Flexible and Adaptive Model-serving for Multi-tenant Edge AI*. Oral presentation delivered at the 8th ACM/IEEE Conference on Internet of Things Design and Implementation (IoTDI), San Antonio, TX, May 11, 2023.
- *AI on the Edge: Rethinking AI-based IoT Applications Using Specialized Edge Architectures*. Oral presentation delivered at IEEE International Symposium on Workload Characterization (IISWC), Otc 29, 2020.

TEACHING EXPERIENCE

University of Massachusetts Amherst – Teaching Assistant

Graduate Independent Study — Spring 2023

- Mentored graduate students, guiding them in identifying research problems, facilitating consistent progress, and successfully completing their independent studies

Reasoning Under Uncertainty (COMPSCI 240) — Spring 2020

Introduction to Informatics (INFO 101) — Fall 2018

- Offered regular office hour, collaborated with the instructor to develop course materials, managed online learning materials, and grading

AWARDS

- The Evan Pugh Scholar Award, The Pennsylvania State University — 2015
  For undergraduate juniors and seniors who are in the upper 0.5 percent of their respective classes.
- The President Sparks Award, The Pennsylvania State University — 2014
  For earning a 4.00(A) cumulative grade point average based on at least 36 graded credits.
- The President’s Freshman Award, The Pennsylvania State University — 2013
  For earning a 4.00(A) cumulative grade point average based on at least 12 graded credits.

SKILLS

- Programming Languages: Python, C/C++, Golang, Rust, JavaScript, HTML5, Bash, LaTeX
- Machine Learning Frameworks: TensorFlow, PyTorch, TensorRT, Nvidia Triton
- Data Science Frameworks: Numpy, Pandas, Scikit-learn
- Operating Systems: UNIX/Linux, OS X, Windows
- Cloud Computing Tools: Docker, Kubernetes, AWS, Ansible