Qiang Xu

+1 (401) 396-6864 | xu1201@purdue.edu | qiangxu1996.github.io | West Lafayette, IN 47907

I enjoy building systems that unlock the full potential of the underlying hardware resources. I am experienced in the development, profiling, and optimization of **machine learning inference systems** for both distributed GPU servers and computationally constrained mobile devices.

Education

Purdue University *West Lafayette, USA*
Ph.D. in Electrical and Computer Engineering *2018 – 2024*
Advisor: Prof. Y. Charlie Hu

University of Science and Technology of China (USTC) *Hefei, China*
B.E. in Computer Science and Technology *2014 – 2018*
Advisor: Prof. Yu Zhang
Talent Program in Computer Science and Technology, School of the Gifted Young

Skills

Programming Python, C/C++, Java, Bash, JavaScript, MATLAB, Julia, SQL, C#
Platforms Linux, CUDA (TensorRT, Nsight), Android, Docker, HPC, Cloud computing
Frameworks PyTorch, ONNX, vLLM, Mobile DL frameworks (ncnn, TensorFlow Lite), RL frameworks (Ray, Gym)
Tools Git, Build systems (CMake, Gradle), gdb, Linux perf, OpenCV, Protobuf, Code coverage (JaCoCo)

Professional Experience

Purdue University *West Lafayette, USA*
Graduate Research Assistant *Aug. 2018 – Present*
Advisor: Prof. Y. Charlie Hu

- Designed a machine learning inference framework for heterogeneous GPU clusters exploiting pipeline parallelism, improving GPU utilization and inference throughput by up to 52.8%. (In preprint)
- Developed a machine-learning-as-a-service framework serving augmented reality clients. The framework maximizes the capacity of a GPU server and supports 1.7x–6.9x more clients concurrently. (**MobiSys 2024**)
- Built the first scheduling framework for augmented reality clients that require offloading multiple machine learning tasks. Improved the overall accuracy by 7.6%–14.3%. (**MobiCom 2023**)
- Characterized the performance of offloading object detection tasks over 5G mmWave in the wild in collaboration with wireless networking research teams. (**MASCOTS 2023**, **5G-MeMU 2022**)
- Surveyed 25 mobile app developers for their practices on deep parameters and energy optimization. Systematically studied and categorized the energy impact of deep parameters in 16 Android apps. (**SANER 2022**)
- Contributed to the design of an energy-aware adaptive bitrate algorithm for video streaming. (**USENIX ATC 2021**)

NEC Laboratories America, Inc. *Princeton, USA*
Research Intern *May 2023 – Aug. 2023*
Mentor: Murugan Sankaradas

- Designed an offloading scheduler to coordinate DNN-powered video analytics clients under network contention. Reduced the request drop rate by up to 92.9% and improved application responsiveness.

Publications

1. **IPIPE: Efficient Video Analytics Serving on Heterogeneous GPU Clusters via Pool-Based Pipeline Parallelism**
   Z. Jonny Kong*, **Qiang Xu***, and Y. Charlie Hu (* co-primary)
   Under submission

2. **ARISE: An Accuracy-Aware Proactive Framework for Serving Concurrent Edge-Assisted AR Clients**
   Z. Jonny Kong*, **Qiang Xu***, and Y. Charlie Hu (* co-primary)
   The 22nd ACM International Conference on Mobile Systems, Applications, and Services (**MobiSys 2024**)

3. **Can 5G mmWave Enable Edge-Assisted Real-Time Object Detection for Augmented Reality?**
   Moinak Ghoshal*, Z. Jonny Kong*, **Qiang Xu***, Zixiao Lu, Shivang Aggarwal, Imran Khan, Jiayi Meng, Yuanjie Li, Y. Charlie Hu, and Dimitrios Koutsonikolas (* co-primary)
   31st International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (**MASCOTS 2023**)

4. **AccuMO: Accuracy-Centric Multitask Offloading in Edge-Assisted Mobile Augmented Reality**
   Z. Jonny Kong*, **Qiang Xu***, and Y. Charlie Hu (* co-primary)
   The 29th Annual International Conference on Mobile Computing and Networking (**MobiCom 2023**)

5. **An In-Depth Study of Uplink Performance of 5G mmWave Networks**
   Moinak Ghoshal*, Z. Jonny Kong*, **Qiang Xu***, Zixiao Lu, Shivang Aggarwal, Imran Khan, Yuanjie Li, Y. Charlie Hu, and Dimitrios Koutsonikolas (* co-primary)
   The 2nd ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Case (**5G-MeMU 2022**)

6. **Can 5G mmWave Support Multi-user AR Apps?**
   Moinak Ghoshal, Pranab Dash, Z. Jonny Kong, **Qiang Xu**, Y. Charlie Hu, Dimitrios Koutsonikolas, and Yuanjie Li
   Passive and Active Measurement Conference 2022 (**PAM 2022**)

7. **An Empirical Study on the Impact of Deep Parameters on Mobile App Energy Usage**
   Qiang Xu, James C. Davis, Y. Charlie Hu, and Abhilash Jindal
   The 29th IEEE International Conference on Software Analysis, Evolution and Reengineering (**SANER 2022**)

8. **Do Larger (More Accurate) Deep Neural Network Models Help in Edge-assisted Augmented Reality?**
   Jiayi Meng, Z. Jonny Kong, **Qiang Xu**, and Y. Charlie Hu
   ACM SIGCOMM 2021 Workshop on Network-Application Integration (**NAI 2021**)

9. **Proactive Energy-Aware Adaptive Video Streaming on Mobile Devices**
   Jiayi Meng, **Qiang Xu**, and Y. Charlie Hu
   2021 USENIX Annual Technical Conference (**USENIX ATC 2021**)

Teaching

Teaching Assistant, Introduction to Operating Systems (ECE 695), Purdue University *2020 – 2023*

Honors and Awards

Ross Fellowship, Purdue University *2018*
National Scholarship (top 0.2% nationwide), USTC *2016*
Outstanding Student Scholarship, USTC *2015*
Outstanding Freshman Scholarship, USTC *2014*