

Z. JONNY KONG

+1(310) 498-9627 ♦ kong102@purdue.edu ♦ www.jonnykong.com

Experienced researcher and engineer specializing in **networked systems, mobile systems, and ML systems**. Seeking opportunities in industry, particularly in roles like Research Scientist, Research Engineer, or Software Engineer.

EDUCATION

Purdue University Ph.D. in Electrical and Computer Engineering	<i>West Lafayette, IN, U.S.</i> <i>Aug 2020 - Present</i>
University of California, Los Angeles M.S. in Computer Science	<i>Los Angeles, CA, U.S.</i> <i>Sep 2018 - June 2020</i>
Beihang University B.E. in Automation	<i>Beijing, China</i> <i>Sep 2014 - June 2018</i>

SKILLS

Programming	Python, C/C++, Java, Bash, Julia, SQL, Lua
Platforms	Linux, CUDA (TensorRT, NSight Systems), Android, Docker, Cloud computing
Frameworks	PyTorch, ONNX, Mobile DL frameworks (TF-Lite, ncnn)
Tools	Git, build systems (CMake, Gradle, buck2), gdb, Linux perf, OpenCV, Protobuf, Thrift

RESEARCH AND PROFESSIONAL EXPERIENCE

Purdue University Research Assistant Advisor: Prof. Y. Charlie Hu	<i>West Lafayette, IN, U.S.</i> <i>Aug 2020 - Present</i>
--	--

- Designed an energy-efficient LLM inference framework on vLLM, leveraging GPU frequency tuning (DVFS) to reduce energy consumption by 18% while maintaining request throughput vLLM, that employs GPU frequency tuning (DVFS) to serve the same number of requests with 18% lower energy
- Designed a machine-learning-as-a-service (MLaaS) framework for GPU clusters using pipelined parallelism, improving serving throughput by up to 52.8% on the same number of GPUs over the industry standard, reducing MLaaS operator's capital expenditure and operating expenses [\[1\]](#)
- Designed an MLaaS framework specifically for serving edge-assisted AR mobile apps, that maximizes the capacities of GPU servers and serves 1.7-6.9x more clients [\[3\]](#)
- Designed an MLaaS framework that optimize the overall accuracy of an AR mobile app that offloads multiple tasks to an edge GPU server, improving the overall accuracy by 7.6%-14.3%, resulting in smoother user experiences [\[5\]](#)
- Conducted measurement studies on latest wireless networks, such as 5G and 802.11ad, in terms of network throughput, handover behaviors, and application performance, revealing their real world performance characteristics [\[7\]](#) [\[4\]](#)

Meta Platforms Systems & Infra Software Engineering Intern	<i>Sunnyvale, CA, U.S.</i> <i>May 2024 - Aug 2024</i>
--	--

- Contributed to the development of IPNext, Meta's latest-generation control plane framework for ads recommendation ML models, using public tools (e.g. C++, Thrift, folly) and Meta-internal tools (e.g. buck2, Thrift, Sapling, JellyFish)
- Implemented a new configuration format for IPNext to streamline the deployment of ads models, reducing configuration file changes per model from three to two, thereby minimizing misconfiguration risks
- Developed verification and rollback procedures, to ensure model migrations to the new configuration will be done correctly and reliably, using tools such as Configurator, Conveyor, Tupperware, Scuba, Thrift Fiddle

- Conducted the migration of all Meta's ads ML models (~100 models deployed globally across 20 regions on 3-4K instances, serving both production ads tasks and periodic offline tasks) without downtime or revenue loss

NOTABLE PUBLICATIONS

- [1] **Z. Jonny Kong***, Qiang Xu*, Y. Charlie Hu. "*IPIPE: Efficient Video Analytics Serving on Heterogeneous GPU Clusters via Pool-Based Pipeline Parallelism*". Under submission. (* co-primary)
- [2] **Z. Jonny Kong**, Nathan Hu, Y. Charlie Hu, Jiayi Meng, Yaron Koral. "*High-Fidelity Cellular Network Control-Plane Traffic Generation without Domain Knowledge*". In **ACM IMC 2024**.
- [3] **Z. Jonny Kong***, Qiang Xu*, Y. Charlie Hu. "*ARISE: An Accuracy-Aware Proactive Framework for Serving Concurrent Edge-Assisted AR Clients*". In **ACM MobiSys 2024**. (* co-primary)
- [4] Moinak Ghoshal*, Imran Khan*, **Z. Jonny Kong***, Phuc Dinh, Jiayi Meng, Y. Charlie Hu, Dimitrios Koutsonikolas. "*Performance of Cellular Networks on the Wheels*". In **ACM IMC 2023**. (* co-primary)
- [5] **Z. Jonny Kong***, Qiang Xu*, Jiayi Meng, Y. Charlie Hu. "*AccuMO: Accuracy-Centric Multitask Offloading in Edge-Assisted Mobile Augmented Reality*". In **ACM MobiCom 2023**. (*co-primary)
- [6] Moinak Ghoshal*, **Z. Jonny Kong***, Qiang Xu*, Zixiao Lu, Shivang Aggarwal, Imran Khan, Jiayi Meng, Yuanjie Li, Y. Charlie Hu, Dimitrios Koutsonikolas. "*Can 5G mmWave Enable Edge-Assisted Real-Time Object Detection for Augmented Reality?*". In **IEEE MASCOTS 2023**. (*co-primary)
- [7] Shivang Aggarwal, **Zhaoning Kong**, Moinak Ghoshal, Y. Charlie Hu, Dimitrios Koutsonikolas. "*Throughput Prediction on 60 GHz Mobile Devices for High-Bandwidth, Latency-Sensitive Applications*". In **PAM 2021 (Best Dataset Award)**. [\[pdf\]](#)

NOTABLE AWARDS

- Best Paper Award, EdgeSys '22
- Best Dataset Award, PAM '21
- National Scholarship of China, 2017 (Top 0.2% nationwide)

PROFESSIONAL SERVICES

Journal Reviewers: IEEE Transactions on Networking, IEEE Transactions on Mobile Computing
Artifact Evaluation Committee (AEC): ACM MobiSys 2023, SOSP 2023

TEACHING ASSISTANT

ECE 26400 Advanced C Programming, Fall '20, Spring '21, Summer '21, Purdue University
CS 151B Computer Systems Architecture, Winter '20, UCLA
CS 217A Internet Architecture and Protocols, Fall '19, UCLA