

# Z. JONNY KONG

+1(310) 498-9627 ♦ [kong102@purdue.edu](mailto:kong102@purdue.edu) ♦ [www.jonnykong.com](http://www.jonnykong.com)

Researcher and engineer specializing in **ML systems, and networked & mobile systems**, with expertise in architecting **scalable infrastructure for real-time AI applications**. Seeking industry roles as a Research Scientist, Research Engineer, or Software Engineer to drive innovation in ML inference and systems optimization.

## EDUCATION

### Purdue University

Ph.D. in Electrical and Computer Engineering

West Lafayette, IN, U.S.

Aug 2020 - Present

### University of California, Los Angeles

M.S. in Computer Science

Los Angeles, CA, U.S.

Sep 2018 - June 2020

### Beihang University

B.E. in Automation

Beijing, China

Sep 2014 - June 2018

## SKILLS

<b>Programming</b>	Python, C/C++, Java, Bash, Julia, SQL, Lua
<b>Platforms</b>	Linux, CUDA (TensorRT, NSight Systems, NVML), Android, Docker, Cloud Platforms (GCP, AWS)
<b>Frameworks</b>	DL (PyTorch, ONNX), LLM (vLLM, DeepSpeed), Mobile DL (TF-Lite, ncnn)
<b>Tools</b>	Git, build systems (CMake, Gradle, buck2), gdb, Linux perf, Wireshark, OpenCV, Protobuf, Thrift

## RESEARCH AND PROFESSIONAL EXPERIENCE

### Purdue University

Research Assistant

West Lafayette, IN, U.S.

Aug 2020 - Present

Advisor: Prof. [Y. Charlie Hu](#)

- Developed an energy-efficient LLM inference framework using GPU frequency tuning (DVFS) on top of vLLM, reducing energy consumption by 18% while preserving request throughput, optimizing cloud inference costs
- Designed a machine-learning-as-a-service (MLaaS) framework for GPU clusters using pipelined parallelism, improving serving throughput by up to 52.8% over the industry standard, reducing MLaaS operator's capital expenditure and operating expenses [\[2\]](#)
- 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 [\[4\]](#)
- 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 [\[6\]](#)
- 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 [\[8\]](#) [\[5\]](#)

### Meta Platforms

Systems & Infra Software Engineering Intern

Sunnyvale, CA, U.S.

May 2024 - Aug 2024

- 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 seamless migration of all Meta ads ML models ( $\approx 100$ ) across 20 regions and 3-4K instances, ensuring zero downtime or revenue impact, streamlining model deployment workflows and improving deployment efficiency

## NOTABLE PUBLICATIONS

---

- [1] **Z. Jonny Kong\***, Omar Basit\*, Abhiram Datla, Abhiram Datla, Y. Charlie Hu. “*Fine-Grained Energy Optimization in LLM Inference via GPU Frequency Control*”. In submission. (\* co-primary)
- [2] **Z. Jonny Kong\***, Qiang Xu\*, Y. Charlie Hu. “*IPIPE: Efficient Video Analytics Serving on Heterogeneous GPU Clusters via Pool-Based Pipeline Parallelism*”. In **USENIX ATC 2025**.
- [3] **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**.
- [4] **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)
- [5] 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)
- [6] **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)
- [7] 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)
- [8] 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)**.
- [9] Jiayi Meng, **Z. Jonny Kong**, Y. Charlie Hu, Mun Gi Choi, Dhananjay Lal. “*Do We Need Sophisticated System Design for Edge-assisted Augmented Reality?*”. In **ACM EdgeSys 2022 (Best Paper Award)**.

## 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 EXPERIENCE

---

**ECE 26400 Advanced C Programming**, Purdue University: TA for Fall '20, Spring '21, Summer '21

**CS 151B Computer Systems Architecture**, UCLA: TA for Winter '20

**CS 217A Internet Architecture and Protocols**, UCLA: TA for Fall '19