Yifei Xu

+1 (310) 728-5031 | yxu@cs.ucla.edu | https://www.yifeix.com

Education

University of California, Los Angeles **Sept. 2021 – Present**
Ph.D. Student in Computer Science *Los Angeles, CA*
Advisor: *Prof. Songwu Lu* ***Research Interests:*** *network systems, ML for systems, wireless networking*

Peking University **Sept. 2017 – July 2021**
B.S. in Computer Science *Beijing, China*

Experience

Alibaba Cloud **June 2022 – Sept 2022**
Research Intern, AIS Networking Research Group **June 2023 – Sept 2023**
Mentor: *Dr. Pan Hu, Dr. Yunfei Ma* ***Research Topics:*** *WiFi for VR streaming, LLM for cloud-native apps* *Sunnyvale, CA*

ByteDance **Apr. 2021 – July 2021**
Networking Research & Development Intern *Beijing, China*

Selected Publications *(2 under review + 8 accepted in total, *co-first author)*

1. **Yifei Xu***, Yuning Chen*, Xumiao Zhang*, Xianshang Lin, Pan Hu, Yunfei Ma, Songwu Lu, Wan Du, Z. Morley Mao, Ennan Zhai, Dennis Cai. CloudEval-YAML: A Realistic and Scalable Benchmark for Cloud Configuration Generation. *NeurIPS 2023 Workshop on ML for Systems (To Appear)*
2. Jinghao Zhao, Zhaowei Tan, **Yifei Xu**, Zhehui Zhang, Songwu Lu. SEED: a SIM-based solution to 5G failures. *SIGCOMM 2022*
3. Zhaowei Tan, Jinghao Zhao, Yuanjie Li, **Yifei Xu**, Songwu Lu. Device-Based LTE Latency Reduction at the Application Layer. *NSDI '21*
4. Jizhou Li*, Zikun Li*, **Yifei Xu***, Shiqi Jiang, Tong Yang, Bin Cui, Yafei Dai and Gong Zhang. WavingSketch: An Unbiased and Generic Sketch for Finding Top-k Items in Data Streams. *SIGKDD 2020*

Selected Projects

CloudEval-YAML - benchmark for cloud configuration generation *(NeurIPS 2023 Workshop on MLSys)* **Alibaba Cloud**
- Led a team to collect over 300 YAML problems on cloud apps including Kubernetes, Envoy and Istio and developed their solutions
- Developed a robust benchmark platform that automates the prompting, query, evaluation and scoring for all problems
- Built a server cluster using a master-worker architecture and optimized Docker image caching, speeding up the evaluation by over $20\times$
- Integrated the benchmark with 12 LLMs for a comprehensive evaluation, incorporating multi-sample query and few-shot prompting

Multiplayer VR Streaming WiFi System - *(under review)* **Alibaba Cloud**
- Built an end-to-end Linux-based multiplayer VR streaming system with customizable performance metrics, incorporating open-source VR streaming solution ALVR and OpenXR runtime Monado
- Designed and implemented a multipath WiFi orchestrator that optimized the global QoE with global cross-layer information
- Established a large-scale emulation leveraging Mininet-WiFi and benchmarked our system against SOTA multipath QUIC schedulers, achieving $35\times$ improvement in tail latency, $1.56\times$ in bitrate and $1.86\times$ in QoE

MobileInsight - open-source mobile network analytics tool **UCLA**
- Enhanced decoding and analytical capabilities for LTE/5G by adding support for MAC layer headers and Control Elements
- Developed automatic analyzers for reconstructing device state traces from collected cellular network packets
- Designed a tracking module that infers the location of cellular IoT device from unencrypted signaling messages

LRP - application layer LTE/5G latency reduction *(NSDI '21, TMC)* **UCLA**
- Co-designed an application layer solution for inferring scheduling parameters and latency reduction in LTE/5G networks
- Built Android-based testbed to validate the solution, achieving a reduction in scheduling-induced latency by $23\% \sim 88\%$
- Collected and analyzed LTE/5G cellular traces in handover and video streaming scenarios for dedicated optimization

Scalable RDMA - for large scale data center networks **ByteDance**
- Co-designed a new RDMA protocol targeting extreme scalability and loss tolerance in data center networks
- Investigated Mellanox OFED source code and prototyped a software-based Linux driver for protocol validation

WavingSketch - high-speed data stream sketching algorithm *(SIGKDD 2020)* **Peking University**
- Devised a sketch-based algorithm for unbiased estimation and top-k query in Gbps network data streams with KBs of memory
- Tailored and evaluated the algorithm in finding frequent items, heavy changes, persistent items & super-spreaders, achieving a $4.5\times$ speed-up and up to $10^6\times$ lower error rate

Technical Skills

Programming Languages: C/C++, Python, MATLAB, Shell script, Java, Lisp
Tools & Platforms: LaTeX, Linux, Redis, PyTorch, Tensorflow, Kubernetes, Docker, Mininet, Wireshark, Android, MobileInsight, srsRAN

Awards and Honors

Excellent Research Award of Peking University **2020**