Department of Computer Science
香港城市大學
City University of Hong Kong

COMPUTER SCIENCE COLLOQUIUM

Edge Computing Systems for Real-time AI-Driven Applications

SPEAKER **Dr. Neiwen Ling**

Postdoctoral Associate
Efficient Computing Lab, Yale University

DATE 28 May, 2024 (Tue)
TIME 10:00 AM - 11:00 AM
VENUE G7315, 7th Floor, Green Zone, Yeung Kin Man Academic Building, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong

ABSTRACT

Edge AI is pivotal for enabling low-latency, privacy-preserving, and resource-efficient solutions across a variety of critical applications, including autonomous driving and embodied AI. Developing robust and optimized Edge AI systems is essential to realizing these advancements. In this talk, I will highlight the opportunities and challenges for deploying Edge AI systems and introduce several edge computing systems I have designed for real-time AI-driven applications. I will begin with my research on supporting concurrent DL tasks on a single resource-constrained edge device. First, I will introduce BlastNet, a model inference abstraction that efficiently utilizes both CPU and GPU resources for concurrent DNN inference. Next, I will discuss RT-mDL, a real-time DL framework that combines model compression and scheduling to enable real-time DL tasks. Following this, I will touch on my research on cooperative Edge AI, showcasing a family of cooperative edge systems that support distributed real-time applications through efficient collaboration among different edge nodes. This includes their implementation in a real-world smart roadside infrastructure system deployed on a university campus. I will conclude by outlining future directions, including my vision for developing edge systems to support LLM-powered autonomous agents.

BIOGRAPHY

Dr. Neiwen Ling is currently a Postdoctoral Associate in the Efficient Computing Lab at Yale, working with Prof. Lin Zhong. She completed her Ph.D. at the Chinese University of Hong Kong, where she was advised by Prof. Guoliang Xing. Her research falls in the intersection of Edge Computing, Machine Learning and Real-time System, with the goal of developing edge computing systems for real-time AI-driven applications, such as autonomous driving. Dr. Ling has published papers on ACM/IEEE flagship conferences, including MobiCom, SenSys, IPSN, MobiSys, and IoTDI. She received one Best Paper Award Finalist and one Best Poster Award from the prestigious international conference SenSys. Besides, she has organized the first FMSys workshop at CPS-IoT Week 2024 and served as a reviewer for top ACM/IEEE journals and conferences like TMC, IMWUT/UbiComp, TOSN, and INFOCOM. She is also on the technical program committee for CHASE 2023 and IoTDI 2023 Poster & Demo.

All are welcome!

In case of questions, please contact Prof GUAN Nan at nanguan@cityu.edu.hk, or visit the CS Departmental Seminar Web at https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums.