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

COMPUTER SCIENCE COLLOQUIUM

Memory-Efficient LLM Inference and Beyond

SPEAKER **Prof. XUE Chun Jason**

Professor
Department of Computer Science,
Mohamed bin Zayed University of
Artificial Intelligence (MBZUAI)

DATE 16 Mar, 2026 (Mon)
TIME 10:00 AM - 11:15 AM
VENUE CS Seminar Room, Y6405, 6th Floor, Yellow Zone, Yeung Kin Man Academic Building, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong

ABSTRACT

The deployment of Large Language Models (LLMs) still faces significant challenges due to their extensive memory requirements. This talk will introduce two innovative approaches to address these challenges: Double Compression and FlexInfer. First, Double Compression combines model compression (quantization and pruning) with lossless data compression, achieving a 2.2x compression ratio while maintaining model accuracy within a 1% drop. It optimizes weight distribution and employs adaptive decompression to balance memory usage and inference speed. Secondly, FlexInfer leverages several advanced system techniques such as prefetching and memory locking to maximize memory efficiency and minimize I/O overhead. It achieves up to 12.5x faster inference under memory constraints compared to traditional methods. Together, these solutions enable memory-efficient LLM deployment, bridging the gap between model size and hardware limitations.

BIOGRAPHY

Prof. Chun Jason Xue is currently a professor of computer science at MBZUAI university, Abu Dhabi. His research focuses on memory and storage systems. He is current associate editor for ACM Transactions on Embedded Computing Systems and ACM Transactions on Storage. He is a fellow of IEEE.

All are welcome!

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