THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Department of Computing
電子計算學系

PolyU Academy for Artificial Intelligence
香港理工大學人工智能高等研究院

Technical Challenges in Inference Architectures for Large AI Models

Prof. Weimin ZHENG
Academician of the Chinese Academy of Engineering
Professor
Department of Computer Science and Technology
Tsinghua University
China

Date : 19 September 2025 (Fri)
Time : 11:00 - 12:00
Venue : PQ304
Language : Putonghua

Abstract

This talk introduces the key technical challenges and architectural solutions in LLM inference. Main contents include: **GPU Memory and Bandwidth Challenges**: Using DeepSeek-R1 (671B parameters) as an example, a single device requires at least 671GB memory (FP8 precision), with bandwidth demands of 740GB/s, highlighting GPU architecture limits. **KV-Cache Storage and Transmission Bottlenecks**: Each token may correspond to tens of KB of KV-Cache. At scale, hundreds of terabytes to petabytes must be processed daily, imposing extreme storage and transmission demands. **Mooncake Architecture – Trading Storage for Computation**: Enables KV-Cache reuse, reducing redundant computation and improving efficiency. Applied in Kimi, Alibaba, and Ant Group, it also won the USENIX FAST Best Paper Award. **KTransformers – CPU-GPU Heterogeneous Cooperative Inference**: Allows trillion-parameter models (e.g., DeepSeek-R1) to run on low-end devices (single CPU+GPU). By leveraging host and GPU memory collaboration and separating attention and MoE layers, it lowers deployment thresholds and accelerates AI PC adoption. **Industry Applications and Open-Source Ecosystem**: Mooncake integrates with inference engines such as vLLM and SGLang, supporting NVIDIA and Huawei Ascend. Its efficiency has been validated in enterprise scenarios at iFLYTEK, Alibaba, and others. Through strategies of “trading storage for computation” (Mooncake) and “enhancing computation with storage” (KTransformers), inference costs and hardware needs are greatly reduced, promoting democratization and personalization of large models, marking the transition of AI PCs from concept to reality.

About the Speaker

Prof. Weimin ZHENG is a Professor at the Department of Computer Science and Technology at Tsinghua University and Academician of the Chinese Academy of Engineering. He has long been engaged in research on high-performance computer architecture, parallel algorithms, and systems. He proposed scalable storage system structures and lightweight parallel extension mechanisms, advancing the theory and methods of storage system scalability.

He led the development and successful application of cluster-architecture high-performance computers in China. His work on ultra-large-scale weather forecasting applications on the domestic Sunway TaihuLight system won the ACM Gordon Bell Prize. He has received the National Science and Technology Progress Award (First Prize once, Second Prize twice), the State Technological Invention Award (Second Prize once), the Ho Leung Ho Lee Science and Technology Progress Award, and the inaugural China Storage Lifetime Achievement Award.

WE DRIVE INNOVATION THROUGH SMART COMPUTING