Department of Electrical Engineering
香港城市大學
City University of Hong Kong

CityU Architecture Lab for Arithmetic and Security (CALAS) Seminar Series

Energy efficient LLM on RISC-V Edge Devices via Dynamic Voltage and Frequency Scaling

Andy, Visiting Student, Imperial College London MSc

Abstract:

The deployment of large language models (LLMs) on edge devices faces critical challenges due to constraints in energy consumption, computational capacity, and operating temperature. Addressing these challenges requires co-optimization across software, system, and hardware layers. Dynamic Voltage and Frequency Scaling (DVFS) is a well-established power management solution in mobile SoCs, dynamically adjusting voltage and frequency based on workload demands to reduce power and thermal overhead. However, existing DVFS strategies are not aware of Large Language Model (LLM) inference workloads, which are often treated as black boxes. Fundamentally, LLM inference consists mainly of addition and multiplication operations, and its power demands can drop significantly when processing sparse data. This observation motivates the hypothesis that embedding sparsity-awareness into models could enable more effective DVFS control. Our tentative research will explore this hypothesis from three complementary directions: Software layer: Develop and evaluate optimization techniques to embed sparsity into LLM inference and identify sparsity patterns. System layer: Design a DVFS governor that leverages sparsity patterns for real-time energy and thermal management. Hardware layer: Build an FPGA-based prototype to simulate, validate, and iteratively refine the proposed approach. We anticipate that this work will contribute new cross-layer methodologies to significantly improve the energy efficiency and thermal sustainability of edge-deployed LLMs.

Biography:

Yue Wu (Andy) received his MSc in Applied Mathematics from Imperial College London and is a prospective PhD student to CALAS. He is an Associate Member of the London Mathematical Society and a former tutor with Cambridge International Education. His research interests span optimization problems, Internet of Things (IoT), and edge AI.

22 September, 2025 (Mon); 2pm – 4pm; P1402; https://cityu.zoom.us/j/96742093029