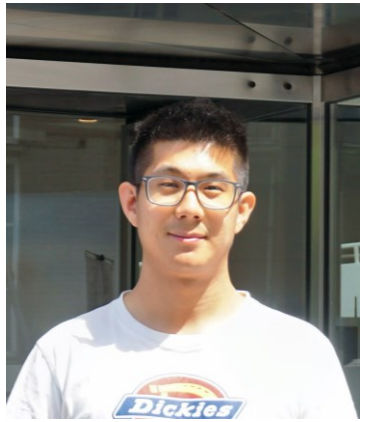
Department of Electrical Engineering
香港城市大學
City University of Hong Kong

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

Accelerating Mamba-Based Vision Models through Algorithm-Hardware Co-Design

Simon Shengzhe LYU, PhD Student

Supervisor: Prof. Ray C.C. CHEUNG and Prof. Weitao XU



Abstract:

State-space models, such as Mamba, offer a compelling alternative to Transformers, reducing quadratic computational complexity while achieving superior accuracy in various tasks, including vision. However, accelerating Vision Mamba (ViM) models faces several challenges, which we address through an efficient algorithm-hardware co-design. First, we propose a hardware-friendly additive power-of-two quantization scheme, applied after smoothing, to mitigate performance degradation in traditional quantization of linear and causal convolutional layers, preserving model accuracy. Additionally, we introduce a unified linear module design that effectively manages quantization and shift-based computations, accommodating the diverse linear layer configurations in ViM. To tackle inefficiencies in state-space updates, which suffer from sequential dependencies and idle times, we develop a fully pipelined scan algorithm to enhance throughput and optimize memory access. Extensive evaluations confirm the effectiveness of these innovations, significantly improving the scalability and practicality of Mamba-based vision models.

Biography:

Simon Shengzhe LYU is a second-year PhD student in CityUHK CS Department. He received his B.Sc. in Electronics Engineering from KU Leuven, BelgiSimonum, and his B.Eng. in Microelectronics Science and Engineering from South China University of Technology, China, both in 2023. Since Sep 2023, he has been pursuing his PhD under the supervision of Prof. Weitao Xu and Prof. Ray C.C. Cheung at CityUHK, focusing on efficient algorithm-hardware co-design for advanced AI algorithms.

18 Aug 2025 (Mon); 2pm - 4pm; G4302; https://cityu.zoom.us/j/96742093029