SEMINAR ANNOUNCEMENT

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
COLLEGE OF DESIGN AND ENGINEERING
Website: https://cde.nus.edu.sg/ece

Area: Integrated Circuits and Embedded Systems (ICES)

Host: Asst Prof Fong Xuanyao Kelvin

TOPIC	:	**Hardware-Software Co-Design of PIM-based LLM Inference Accelerator**
SPEAKER	:	**Mr Chong Yue Jiet** **Graduate Student, ECE Dept, NUS**
DATE	:	**Friday, 10 October 2025**
TIME	:	**11:00AM to 12:00PM**
VENUE	:	**E1-06-08**

ABSTRACT

Both Large-Language Model (LLM) sizes and inference activities have increased exponentially in recent years, causing booming of computing power demands. However, the traditional von Neumann architecture is facing two major bottlenecks, namely memory wall and high-power consumption for AI applications. We present a chiplet-based large language model (LLM) inference accelerator, consisting of non-volatile in-memory-computing processing elements (PEs) and Inter-PE Computational Network (IPCN) to effectively address the communication bottlenecks. Our results show it achieves 3.95× speedup and 30× efficiency improvement over the Nvidia A100.

BIOGRAPHY

Jason Chong is currently a graduate student at ECE, under supervision of Asst. Prof. Fong Xuanyao, Kelvin. His research focuses on hardware-software co-design of AI Accelerator. Prior to joining NUS, he spent two years as SoC-RTL Design Engineer at Intel Corporation

https://cde.nus.edu.sg/ece/highlights/events/