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	:	**Scaling LLMs on Heterogeneous In-Memory Architectures: An Architectural Perspective**
SPEAKER	:	**Mr Abhishek Tyagi** **Graduate Student, ECE Dept, NUS**
DATE	:	**Friday, 7 November 2025**
TIME	:	**1:00PM to 2:00PM**
VENUE	:	**E1-06-08**

ABSTRACT

As Large Language Models (LLMs) continue to grow in scale and capability, the limits of traditional von Neumann architectures have become increasingly evident. The massive memory bandwidth demands and irregular dataflows of transformer-based models expose deep inefficiencies in conventional GPU and CPU systems. This seminar examines whether heterogeneous systems - combining digital and ReRAM-based in-memory computing (IMC) cores - offer a practical and scalable path for LLM deployment. It analyzes how LLM subcomponents such as attention and feed-forward networks differently stress compute and communication, and how architectural strategies for tensor partitioning, workload mapping, and chiplet integration can mitigate these bottlenecks. Emphasizing system-level design the seminar explores the trade-offs between performance, utilization, and scalability, posing a central question for future AI hardware: are heterogeneous IMC-based architectures a viable solution for large-scale language models, or an elegant yet impractical ideal?

BIOGRAPHY

Abhishek is currently a graduate student at ECE, under supervision of Asst. Prof. Kelvin Fong and Assoc. Prof. Bharadwaj Veeravalli. His research focuses on hardware-software co-design of AI Accelerator. Prior to joining NUS, he spent two years as a GPU Architect at NVIDIA.

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