Topic: GPU-Free Computing for LLM Inference

Presenter: Chintan Mehta

Research Questions: How can the memory bandwidth and capacity challenges in LLM inference be addressed without GPUs? Can a GPU-free, memory centric architecture deliver higher throughput, lower latency, and improved energy efficiency compared to modern GPU systems?

Challenge: Modern LLMs feature large parameter sizes and long context windows, resulting in extensive key-value (KV) caches. Due to the sequential nature of LLM decoding, GPUs are underutilized, making inference a memory-bound task. This mismatch necessitates a redesign of hardware for efficient inference.

Claims: The proposed CENT architecture achieves 2.3x higher throughput, consumes 2.3x less energy, and delivers 5.2x more tokens per dollar compared to state-of-the-art GPU systems.

Evidence: Evaluations using Llama2 models show that key metrics such as geomean end-to-end query latency, throughput (tokens/sec) and energy efficiency (tokens/joule), consistently favor CENT over GPU baselines.

Statistical Analysis: Benchmarking across various model sizes (7B, 13B, and 70B) for varying batch sizes and context lengths confirms significant performance gains, validating CENT's design in overcoming LLM inference memory bottlenecks.