MVJ COLLEGE OF ENGINEERING
Since 1982
Engineering A Better Tomorrow

An Autonomous Institute
Approved by AICTE, New Delhi
Affiliated to VTU, Belagavi
Recognized by UGC under 2(f) & 12(B)
Accredited by NBA & NAAC

Add on Lecture - Communication Networks

The Departments of **Electronics and Communication (Advanced Communication Technology) and Industrial IoT** organised an **Add-on Lecture** titled **Communication Networks** on **17-03-2026** at **9:45 AM**, in **Seminar Hall-3**.

The resource person for the event was Dr. Parimal Parag, an Associate Professor in the Department of Electrical Communication Engineering at the Indian Institute of Science (IISc), Bengaluru. A total of 102 students from EC(ACT) and IIoT departments attended the lecture.

Objectives of the Event

The primary objectives of the event were:

- To understand the inference workflow of LLM serving systems, including the roles of the prefill and decode phases and how Key-Value (KV) caching reduces redundant computations.
- To explore performance challenges in LLM decoding, particularly underutilisation of hardware resources, and examine modern optimisation strategies such as alternating prefill–decode operations and prompt prefilling.
- To learn scheduling and control mechanisms—like threshold-based switching and optimal departure structure—that improve throughput, reduce latency, and enhance resource utilisation in large-scale LLM serving systems.

Event Overview

The session commenced at 9:45 AM with a welcome address by Prof. Farha Kowser, Assistant Professor department of EC(ACT), followed by an engaging session by **Dr. Parimal Parag, an Associate Professor in the Department of Electrical Communication Engineering at the Indian Institute of Science (IISc), Bengaluru**.

The presentation covered key areas such as LLM Inference Workflow, Performance Bottlenecks in Decoding, Optimisation Techniques for Efficient Serving, Scheduling, and Control Mechanisms, providing deep insights into the subject matter. Students showed keen interest, with active participation during interactive segments.

MVJ COLLEGE OF ENGINEERING
Since 1982
Engineering A Better Tomorrow

An Autonomous Institute
Approved by AICTE, New Delhi
Affiliated to VTU, Belagavi
Recognized by UGC under 2(f) & 12(B)
Accredited by NBA & NAAC

Figure 1: Prof. Farha Kowser introducing Dr. Parimal Parag

This lecture focused on how Large Language Models (LLMs) are optimised for efficient inference in real-world serving systems. It began with an explanation of the two main stages of LLM inference—prefill (processing the full prompt) and decode (generating tokens one-by-one)—and highlighted the computational challenges associated with these phases. A major technique discussed was the Key-Value (KV) cache, which stores intermediate attention results to avoid re-computation and speed up inference.

Dr. Parimal then addressed the key performance bottlenecks, especially the underutilisation of hardware resources during decoding, which often limits throughput. To overcome this, the lecture explored modern optimisation strategies used in advanced LLM systems such as IBM Granite and LLaMA3. These include prompt prefilling, decoding prefetched prompts, and alternating between prefill and decode operations to maintain steady resource usage.

Finally, the lecture covered scheduling mechanisms—particularly threshold-based switching and the concept of optimal departure structure—that help decide when the system should transition between prefill and decode stages. These strategies collectively aim to improve resource utilisation, reduce latency, and enhance scalability in LLM serving environments.

Figure 2: Dr. Parimal Parag delivering the lecture

Outcomes and Impact

- Understanding LLM Inference Phases: The lecture highlighted how LLMs process queries through two main phases—prefill and decode. Prefill handles the full prompt at once, while decode generates output tokens sequentially, making optimisation crucial.
- Importance of KV Caching: A major takeaway was the role of Key-Value (KV) caching in preventing repeated computation during attention operations, which significantly speeds up inference and reduces computational overhead.
- Decoding Inefficiencies and Optimisation Strategies: The speaker emphasised challenges such as GPU/CPU underutilisation during decoding, and introduced modern optimisation methods used in systems like IBM Granite and LLaMA3—e.g., prefetching prompts, alternating prefill and decode, and pipeline balancing.
- More Scalable and Efficient LLM Serving: The overall objective is to make LLM deployments more efficient, cost-effective, and capable of supporting high volumes of user requests.

Conclusions

The event successfully highlighted how modern LLM serving systems can be optimised for better performance and scalability. Participants gained a clear understanding of the prefill–decode workflow, KV caching, and the scheduling strategies that improve throughput and resource utilisation. The session effectively demonstrated

practical techniques used in real-world models, providing valuable insights into reducing latency and maximising hardware efficiency. Overall, the lecture achieved its goals by enhancing both conceptual clarity and practical awareness of inference optimisation.

Report by: Prof. Sheher Banu S

Department of EC(ACT),

MVJ College of Engineering