,

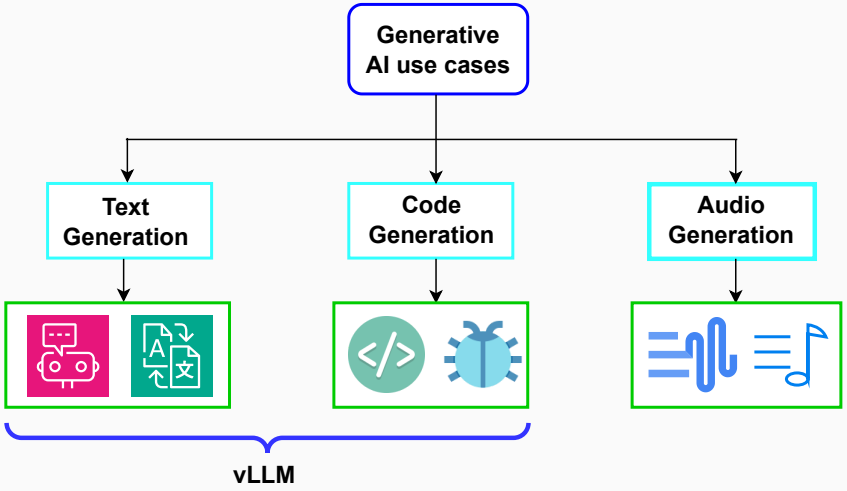
Deferred prefill for throughput maximization in LLM inference

Moonmoon Mohanty

31-Mar-2025

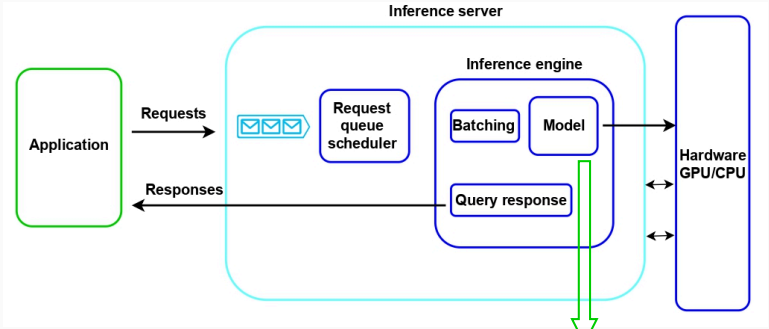
IISc PI: Parimal Parag
Team members (IISc): Gautham Bolar, Preetam Patil
Team members (IBM): UmaMaheswari Devi, Felix George, Pratibha Moogi

Motivation

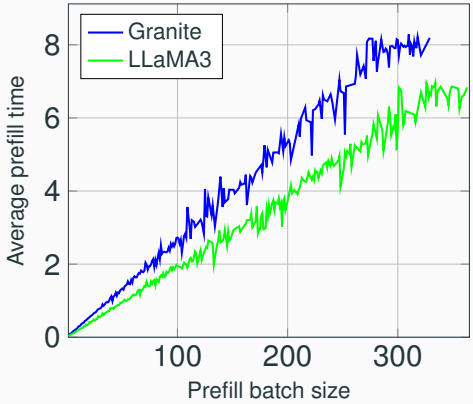


- Cost, energy reduction
- Better user experience

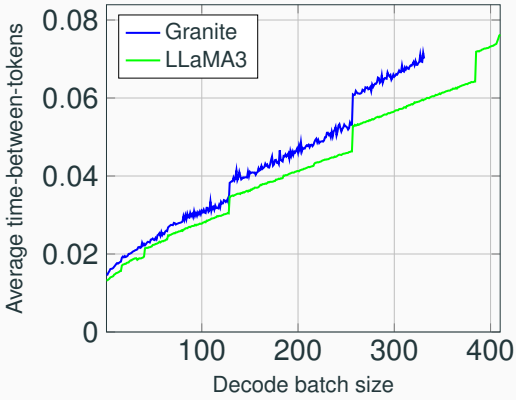
LLM inference



Prefill and Decode times(ShareGPT dataset)



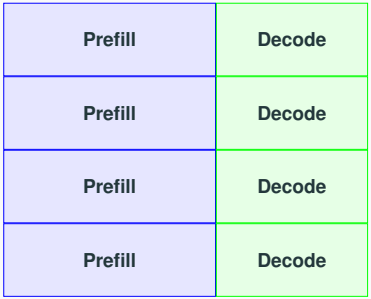
Prefill time



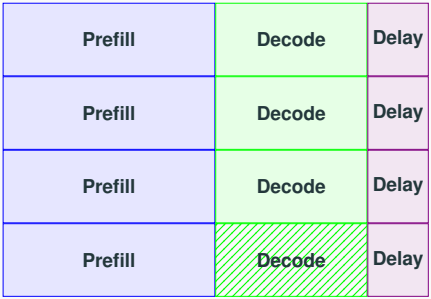
Decode time

Prompt processing policy

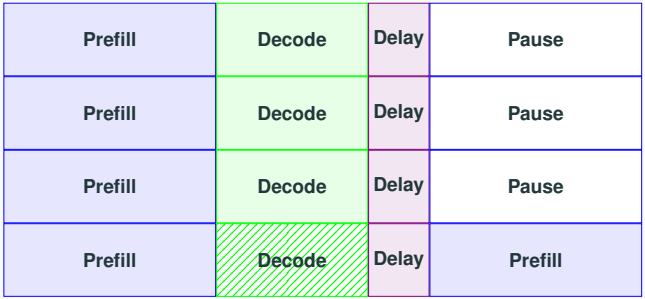
Prompt processing policy



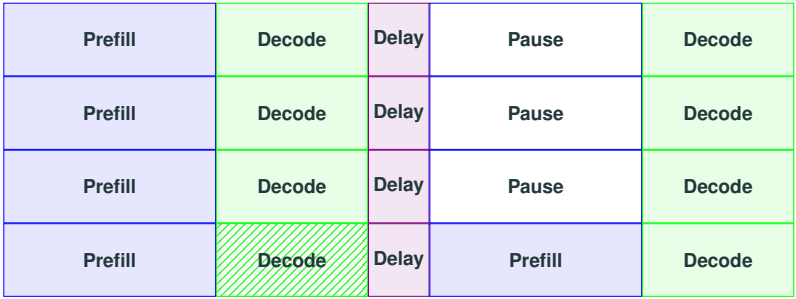
Prompt processing policy



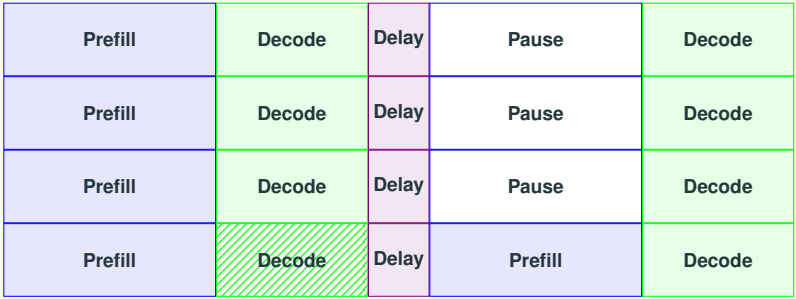
Prompt processing policy



Prompt processing policy



Prompt processing policy



Challenge: Frequent prefills = larger switching delay.
Solution: Prefill after multiple departures

Key question

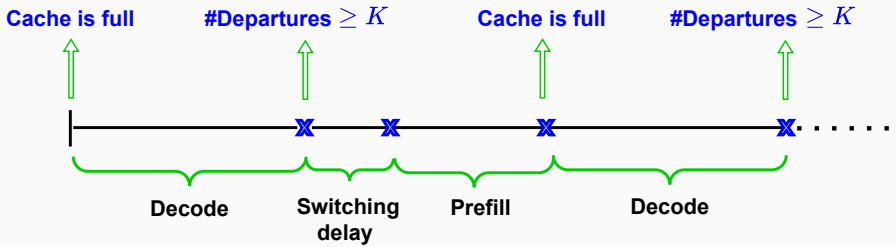


Figure 1: Timeline

Goal: Optimal departure threshold for average throughput maximization?

System Model

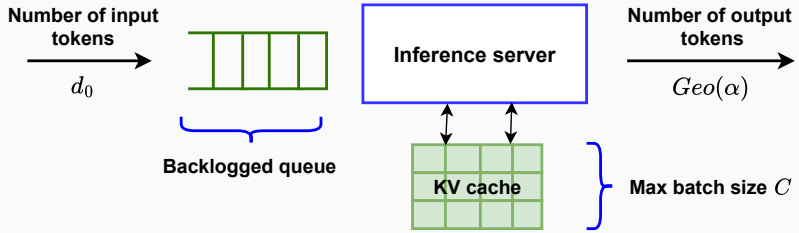
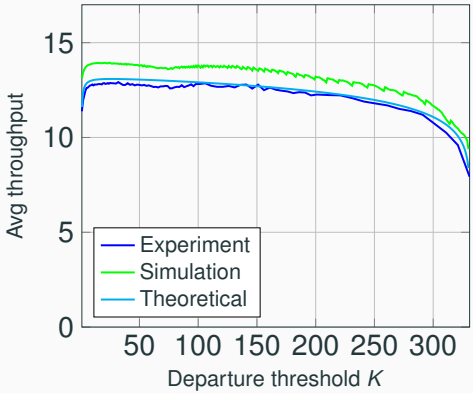


Figure 2: System model

$$\rho(K)^{-1} \approx \frac{1}{K}\left(c_p + c_d \frac{\ln(1-\frac{K}{C})}{\ln(1-\alpha)}\right) + \frac{t_d}{\alpha} + \frac{t_p d_0}{N}$$

K: departure threshold
c_p: scheduling overhead
t_p: i/p processing time
c_d: o/p token compute
t_d: memory slowdown

Experimental validation



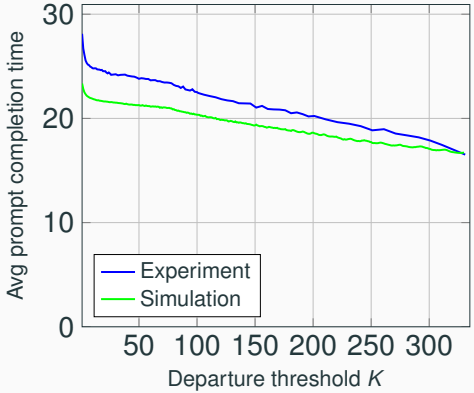
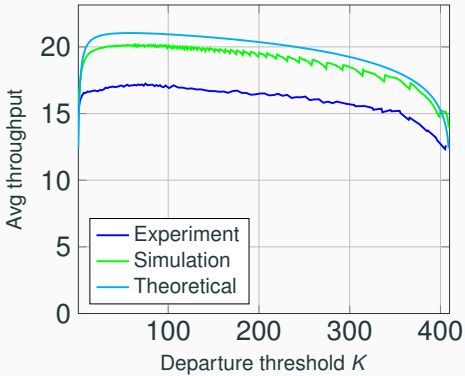


Figure 3: Performance metrics for the Granite model with the ShareGPT dataset.

Experimental validation



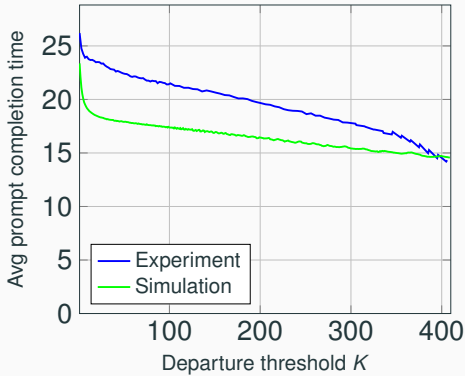


Figure 4: Performance metrics for the LLaMA3 model with the ShareGPT dataset.

Conclusion

- Departure threshold-based scheduling algorithm.
- Analytical model for inference system, deduced closed form expression for throughput.
- Proved existence of optimal departure threshold that maximizes the system throughput.
- Characterization of LLM inference system for system parameters.
- Experimental validation with vLLM inference server and NVIDIA A100 GPU.
- **Key observation**: Proposed policy leads to 13% improvement in average throughput accompanied by 14% reduction in average prompt completion time.