

same machine with NUMA node isolation to further improvement in tokens/s. Table 2, we have received **4x** additional improvement with 4 workers. This would also make Gen-AI based products and companies' environment friendly, our estimates shows that CPU usage for Inference could reduce the power consumption of LLMs by **48.9%** (1252 W for A100 with AMD EPYC 7V13 vs 613 W for Intel® Xeon® Gold 6538N) while providing production ready throughput & latency.

1. Introduction

The widespread integration of large language models (LLMs) across diverse applications, ranging from code generation to writing tasks, has surged in recent times, creating a heightened demand for more efficient inference solutions. Enterprises are increasingly leveraging LLMs to enhance various internal processes. However, the cost associated with their utilization remains a significant concern due to the reliance on GPUs for inference. [1]

The current sequential approach employed by LLMs involves generating one token at a time based on the input prompt and the preceding tokens, persisting until a stop token or the predetermined maximum tokens are reached for each request. This method, while functional, restricts the optimal utilization of available resources [2], [3]. Elevating the tokens generated per second is crucial in mitigating the expenses associated with running LLM-enabled applications. While batching requests can increase throughput, achieving superior batching necessitates optimized memory utilization corresponding to the batch.

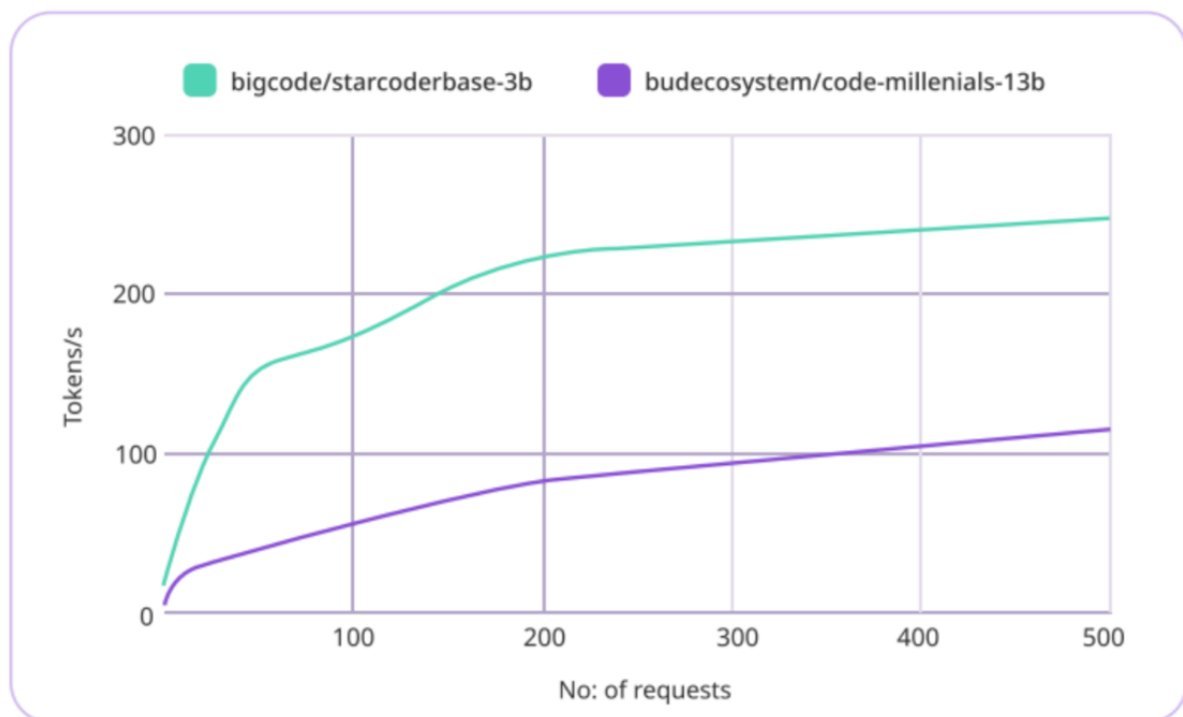


Figure 2: This shows the tokens/s increases with the number of parallel requests increases due to the better utilization of the memory