Benchmarking Multi-LoRA Adapters on the vLLM TPU and GPU Backends Using Llama3.1 models

Saheli Bhattacharjee, Akshat Tripathi, Junfan Huang, Anton Lokhmotov (KRAI, Cambridge)

We present a comprehensive performance analysis of serving multiple LoRA adapters for the Llama3.1-8B and Llama3.1-70B models across the TPU and GPU backends of the vLLM serving framework. We conduct experiments using the MLPerf OpenOrca dataset with up to 2K token contexts, and then extend our evaluation to longer sequences using NVIDIA's GenAI-Perf benchmarking tool to measure critical inference metrics.

Our methodology captures Time to First Token (TTFT), Time Per Output Token (TPOT), and overall request throughput across varying sequence lengths and adapter configurations. The results demonstrate distinct performance characteristics between TPU and GPU backends when serving multiple LoRA adapters concurrently through vLLM's efficient serving architecture.

Benchmarking with the OpenOrca dataset reveals that the TPU backend demonstrates different throughput scaling properties compared to the GPU backend, particularly when serving multiple adapters simultaneously. For the 70B model, the performance differences become especially pronounced in multi-tenant serving environments. As the sequence length increases beyond 2K tokens in the GenAI-Perf evaluation, we observe divergent performance patterns between the hardware platforms.

The benchmark identifies specific crossover points where the preferred hardware backend shifts based on workload characteristics, sequence length, and the number of concurrently served adapters. Memory utilization scaling follows significantly different patterns between backends, with vLLM's paged attention mechanism showing variable efficiency on TPUs versus GPUs when increasing both adapter count and sequence length.

This research provides deployment engineers with concrete performance insights for optimal hardware selection when using vLLM to serve multiple LoRA adapters for Llama 3.1 models across different infrastructure configurations.