Online Result Summary

Model: text_recognition

GPU(s): 2 x NVIDIA A100-SXM-80GB

Total Available GPU Memory: 158.6 GB

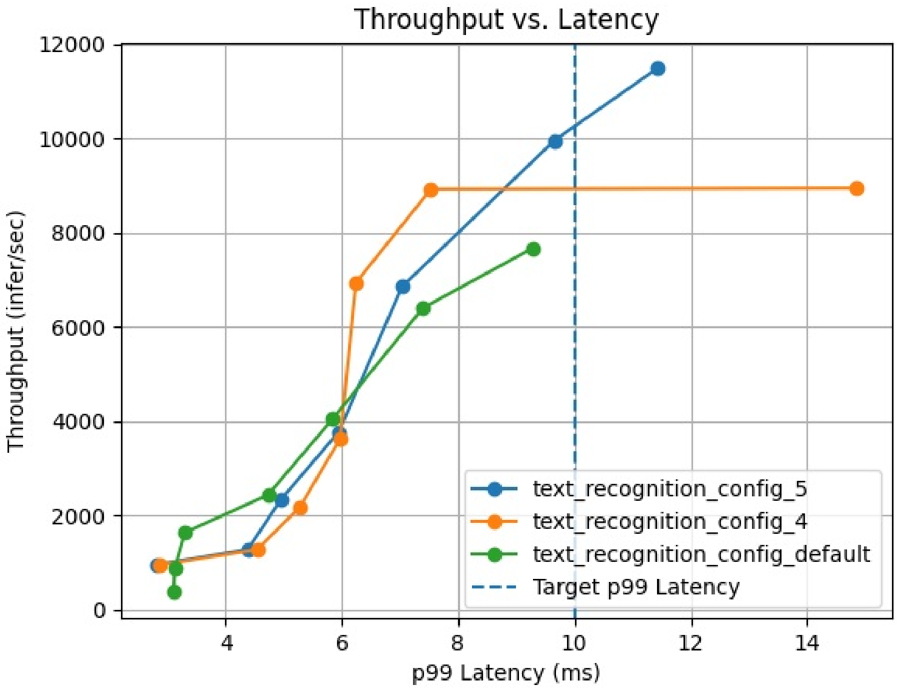
Constraint targets: Max p99 Latency : 10 ms

In 274 measurements across 30 configurations, **text_recognition_config_5** provides the best throughput: **9960 infer/sec**.

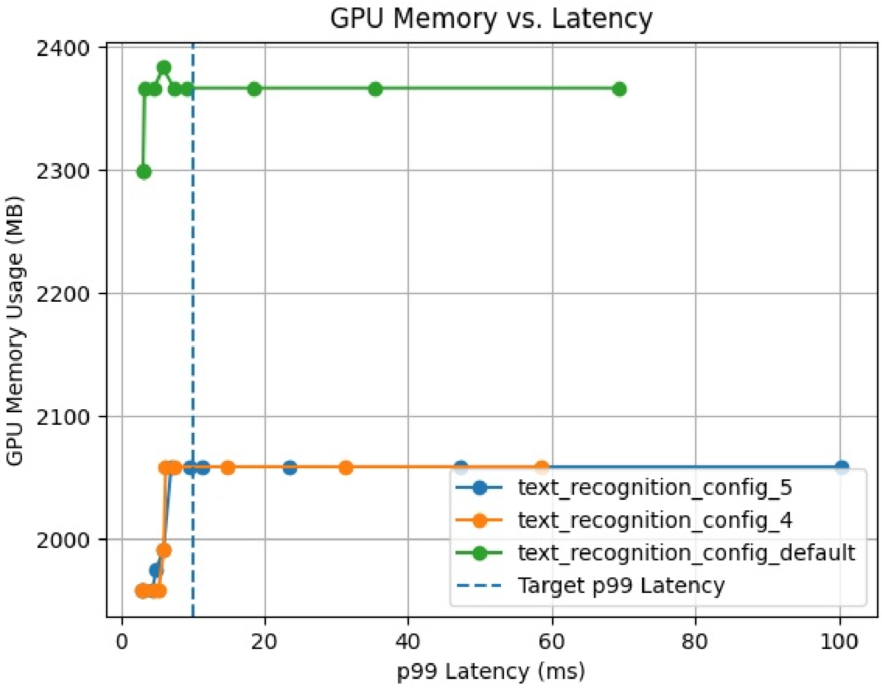
This is a **30% gain** over the default configuration (7671 infer/sec), under the given constraints on GPU(s) 2 x NVIDIA A100-SXM-80GB.

- **text_recognition_config_5**: 2 GPU instances with a max batch size of 32 on platform onnxruntime_onnx

Curves corresponding to the 3 best model configuration(s) out of a total of 30 are shown in the plots.



Throughput vs. Latency curves for 3 best configurations.



GPU Memory vs. Latency curves for 3 best configurations.

The following table summarizes each configuration at the measurement that optimizes the desired metrics under the given constraints.

Model Config Name	**Max Batch Size**	**Dynamic Batching**	**Instance Count**	**p99 Latency (ms)**	**Throughput (infer/sec)**	**Max GPU Memory Usage (MB)**	**Average GPU Utilization (%)**
text_recognition_config_5	32	Enabled	2:GPU	9.658	9959.67	2058	82.6
text_recognition_config_4	16	Enabled	2:GPU	7.532	8925.97	2058	61.4
text_recognition_config_default	8	Enabled	4:GPU	9.279	7670.59	2366	72.0