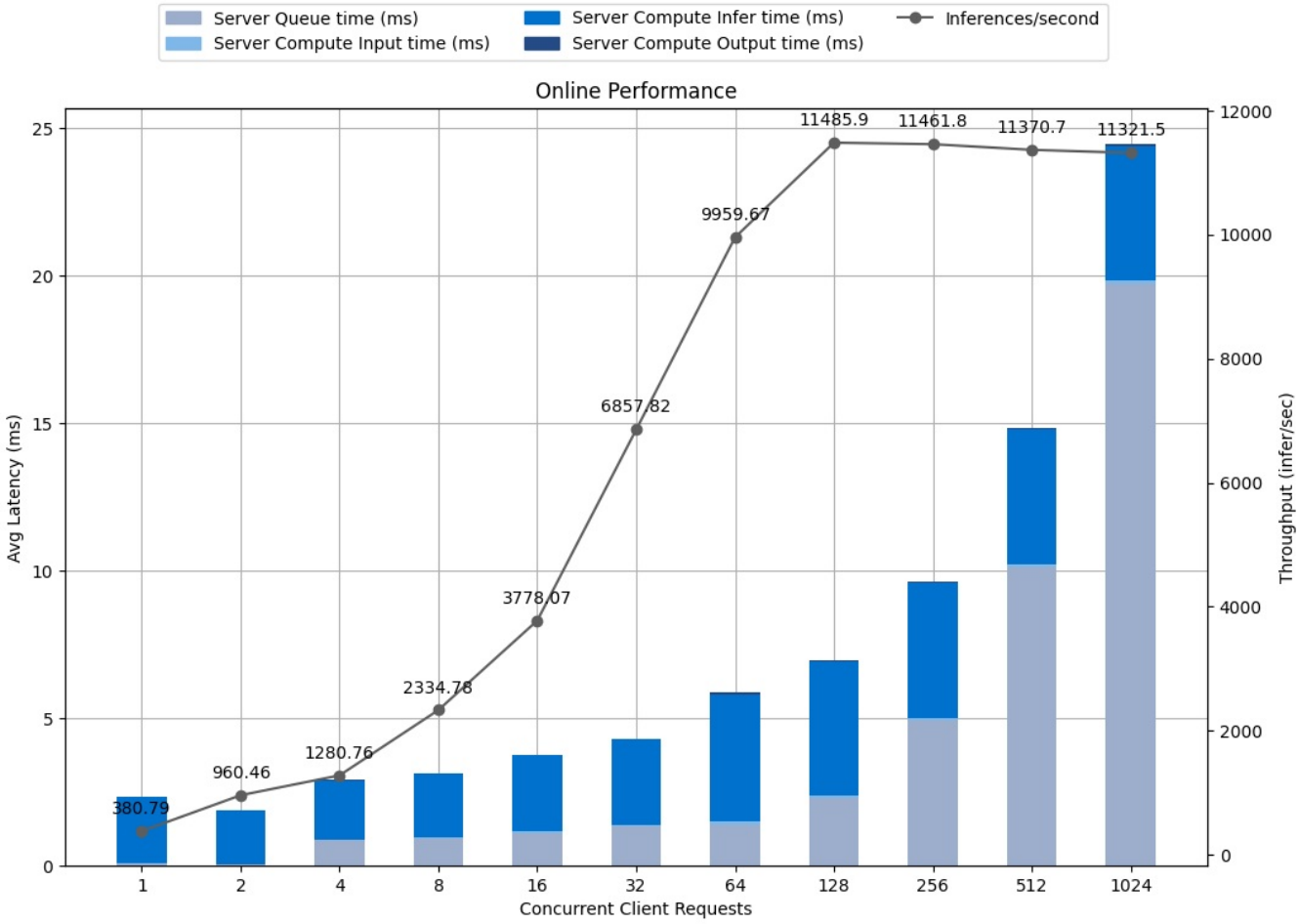
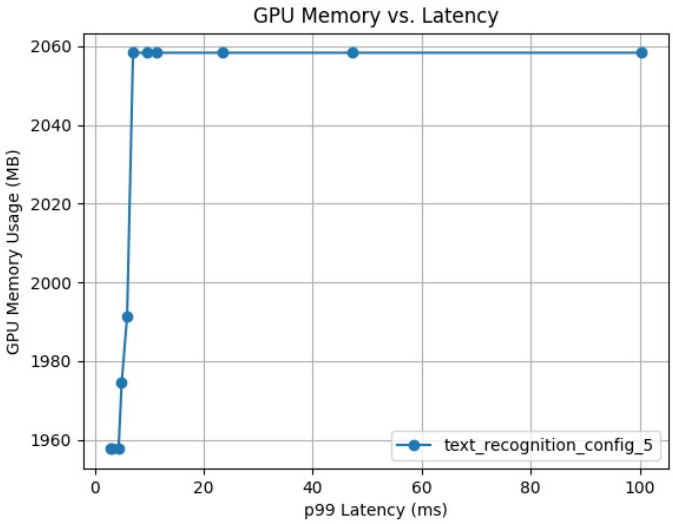
Detailed Report

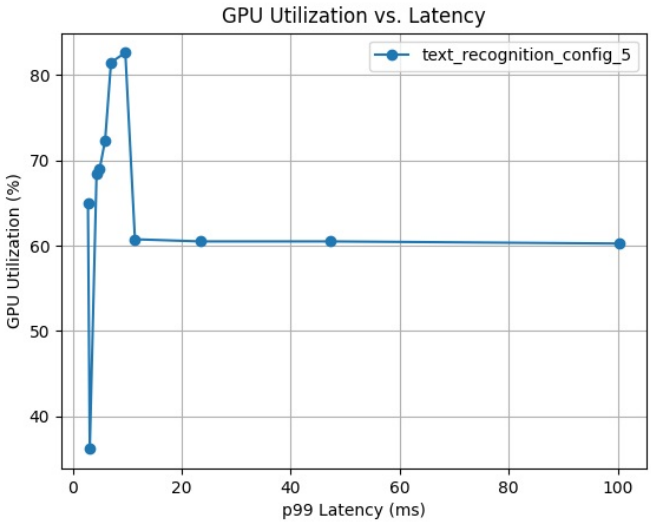
Model Config: text_recognition_config_5



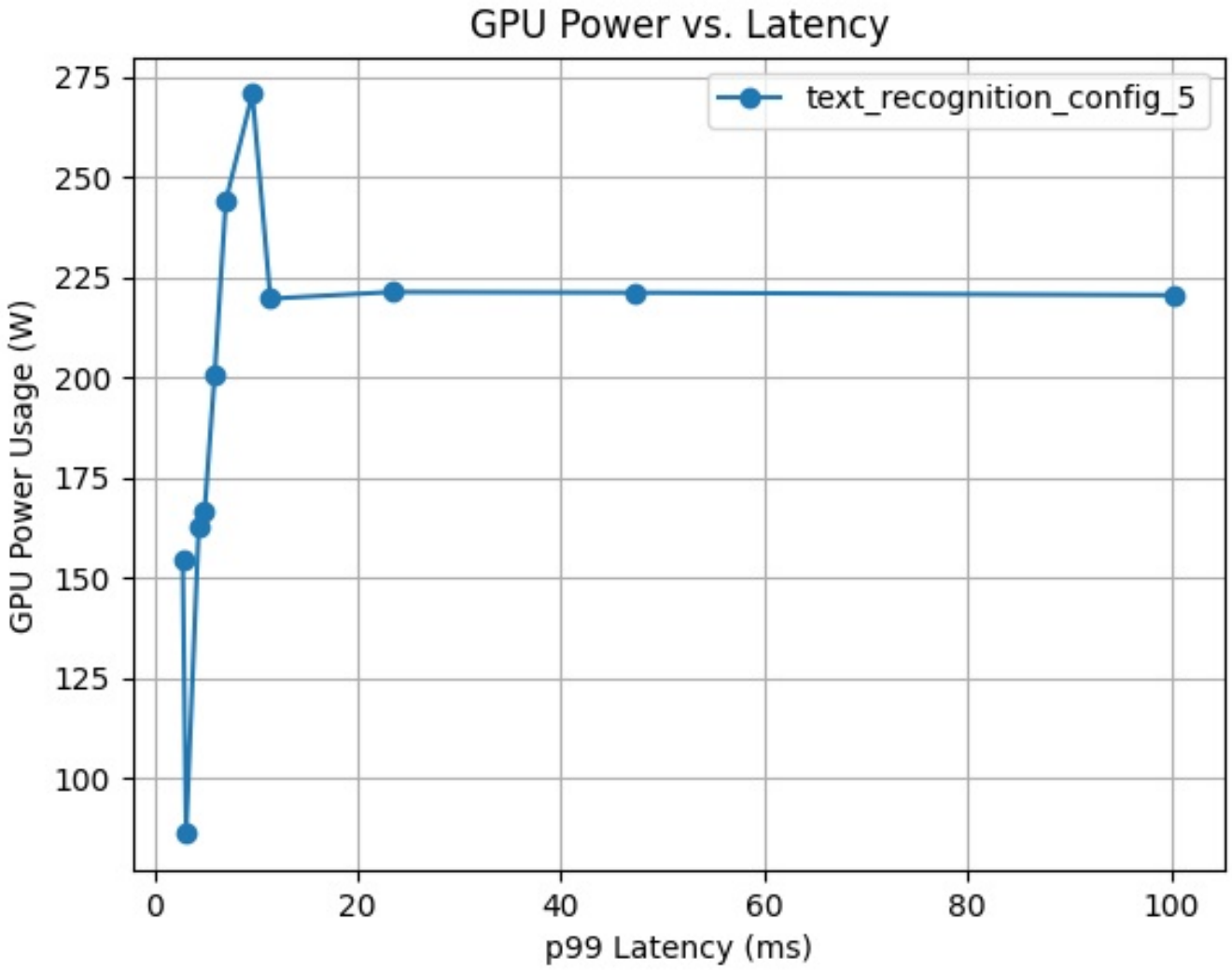
Latency Breakdown for Online Performance of text_recognition_config_5



GPU Memory vs. Latency curves for config text_recognition_config_5



GPU Utilization vs. Latency curves for config text_recognition_config_5



GPU Power vs. Latency curves for config text_recognition_config_5

Request Concurrency	p99 Latency (ms)	Client Response Wait (ms)	Server Queue (ms)	Server Compute Input (ms)	Server Compute Infer (ms)	Throughput (infer/sec)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
1024	100.405	88.671	19.719	0.11	4.577	11321.5	2058.354688	60.2
512	47.331	44.569	10.118	0.105	4.57	11370.7	2058.354688	60.5
256	23.406	22.221	4.918	0.097	4.562	11461.8	2058.354688	60.5
128	11.424	11.105	2.292	0.089	4.548	11485.9	2058.354688	60.8
64	9.658	6.41	1.36	0.121	4.333	9959.67	2058.354688	82.6
32	7.036	4.649	1.33	0.06	2.889	6857.82	2058.354688	81.5
16	5.948	4.219	1.094	0.054	2.587	3778.07	1991.245824	72.3
8	4.948	3.414	0.937	0.031	2.141	2334.78	1974.4686080000001	69.0
4	4.383	3.11	0.86	0.023	2.011	1280.76	1957.691392	68.5
1	3.136	2.611	0.051	0.025	2.269	380.786	1957.691392	36.2
2	2.803	2.072	0.032	0.022	1.835	960.461	1957.691392	65.0

The model config "text_recognition_config_5" uses 2 GPU instances with a max batch size of 32 and has dynamic batching enabled. 11 measurement(s) were obtained for the model config on GPU(s) 2 x NVIDIA A100-SXM-80GB with total memory 158.6 GB. This model uses the platform onnxruntime_onnx.

The first plot above shows the breakdown of the latencies in the latency throughput curve for this model config. Following that are the requested configurable plots showing the relationship between various metrics measured by the Model Analyzer. The above table contains detailed data for each of the measurements taken for this model config in decreasing order of throughput.