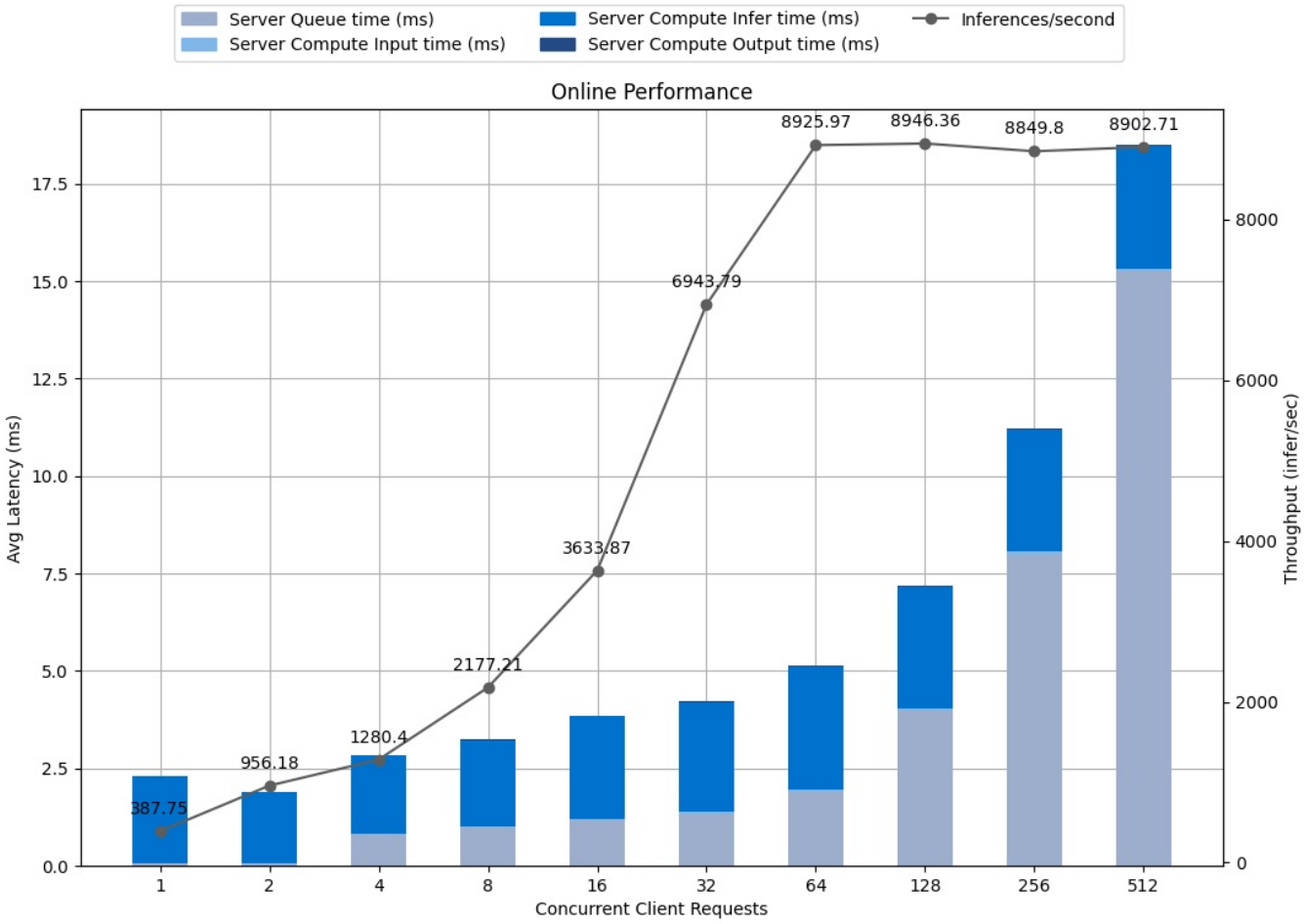
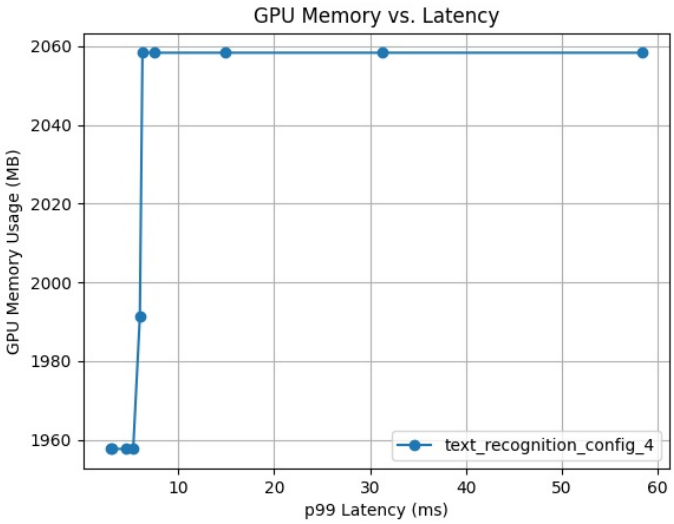
Detailed Report

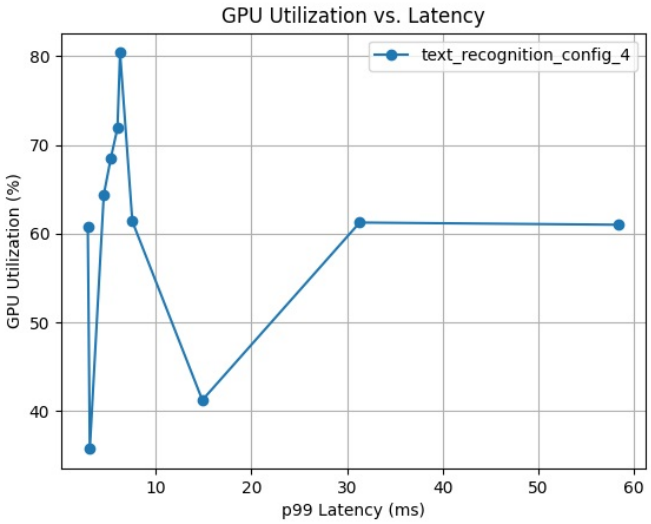
Model Config: text_recognition_config_4



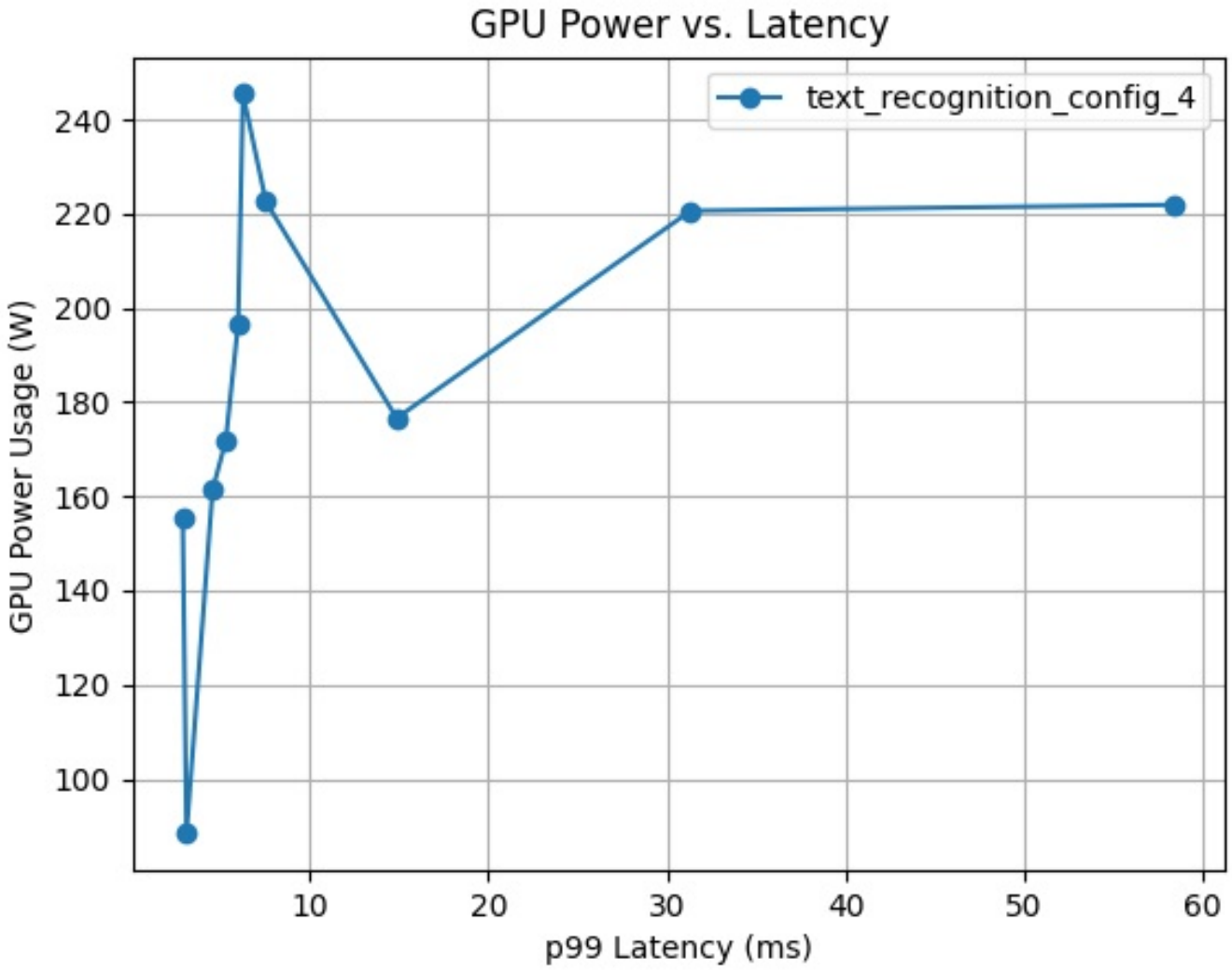
Latency Breakdown for Online Performance of text_recognition_config_4



GPU Memory vs. Latency curves for config text_recognition_config_4



GPU Utilization vs. Latency curves for config text_recognition_config_4



GPU Power vs. Latency curves for config text_recognition_config_4

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 (%)
512	58.481	56.817	15.262	0.07	3.128	8902.71	2058.354688	61.0
256	31.334	28.742	7.989	0.067	3.135	8849.8	2058.354688	61.2
128	14.858	14.253	3.958	0.065	3.144	8946.36	2058.354688	41.2
64	7.532	7.142	1.88	0.062	3.161	8925.97	2058.354688	61.4
32	6.244	4.594	1.325	0.069	2.813	6943.79	2058.354688	80.4
16	5.972	4.383	1.155	0.056	2.628	3633.87	1991.245824	71.9
8	5.276	3.659	0.984	0.032	2.207	2177.21	1957.691392	68.5
4	4.542	3.109	0.788	0.024	2.027	1280.4	1957.691392	64.4
1	3.078	2.564	0.052	0.025	2.222	387.753	1957.691392	35.8
2	2.868	2.082	0.034	0.021	1.835	956.178	1957.691392	60.8

The model config "text_recognition_config_4" uses 2 GPU instances with a max batch size of 16 and has dynamic batching enabled. 10 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.