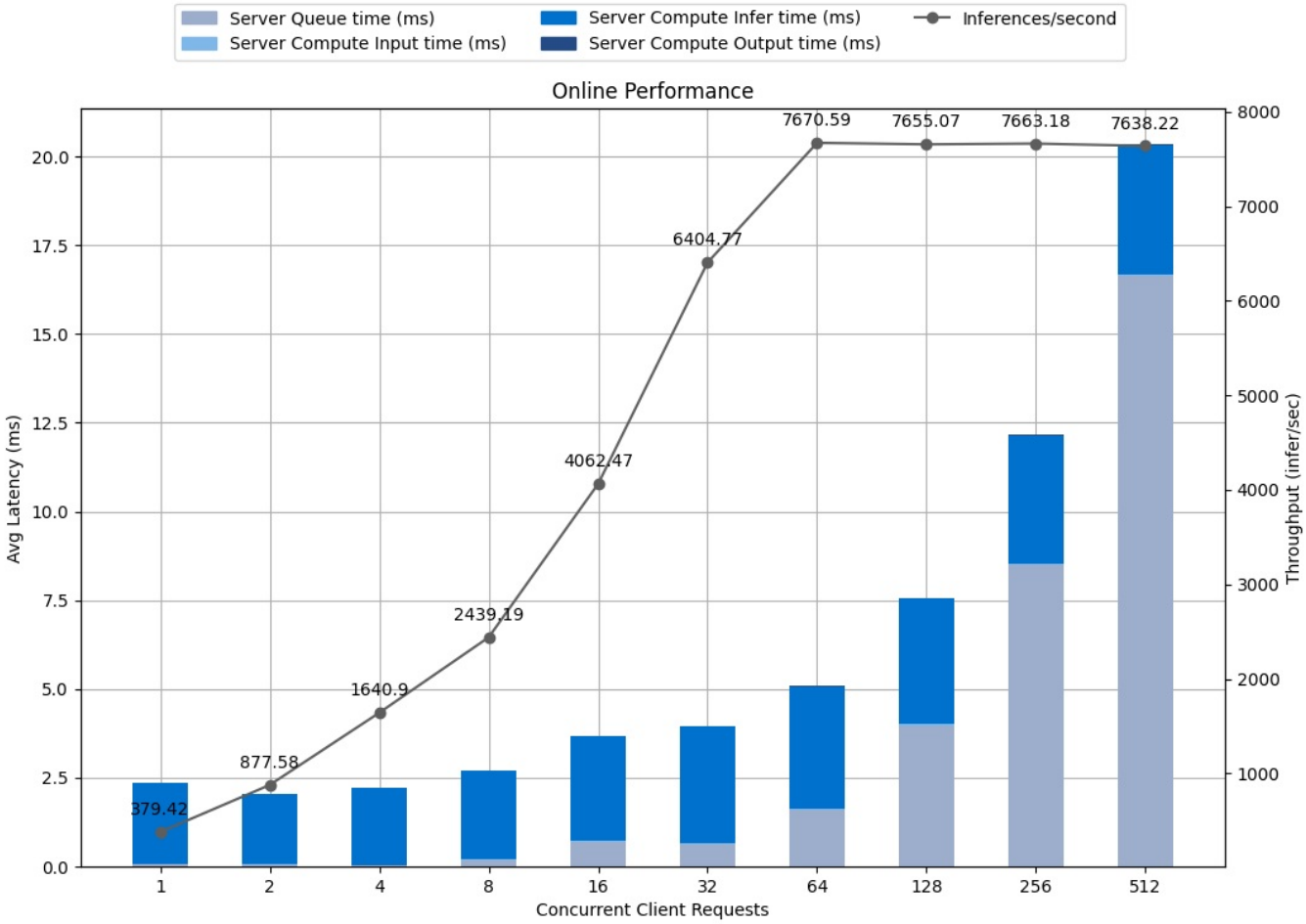
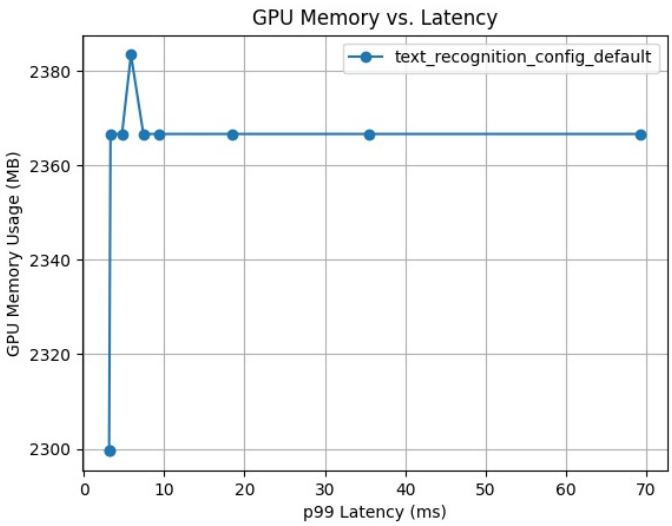
Detailed Report

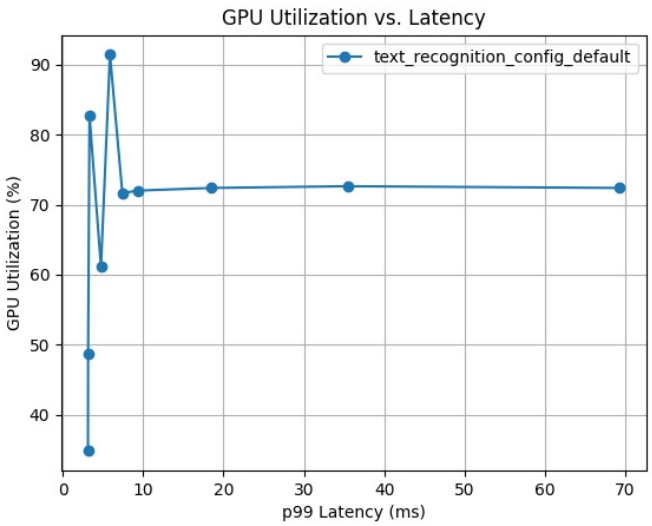
Model Config: text_recognition_config_default



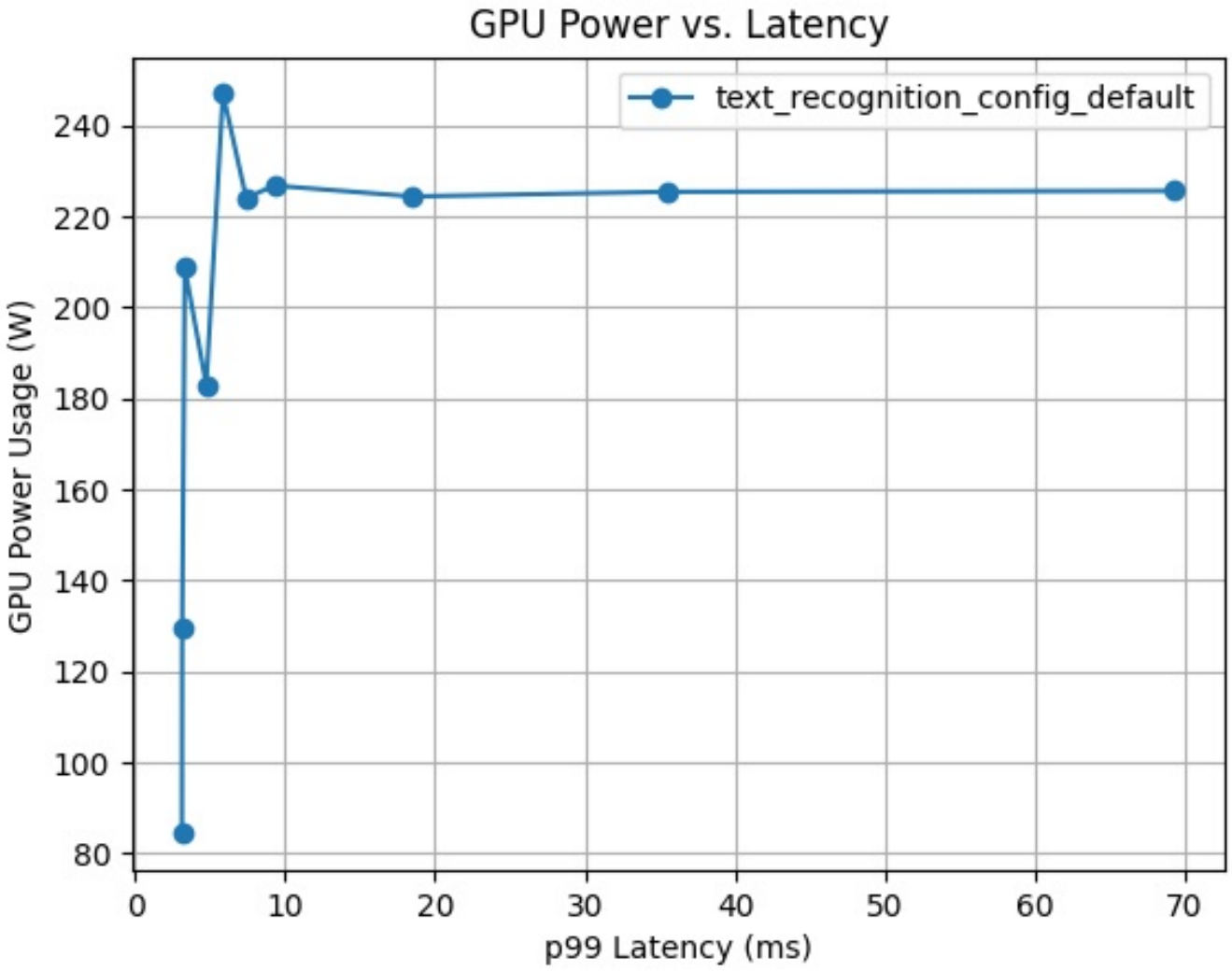
Latency Breakdown for Online Performance of text_recognition_config_default



GPU Memory vs. Latency curves for config text_recognition_config_default



GPU Utilization vs. Latency curves for config text_recognition_config_default



GPU Power vs. Latency curves for config text_recognition_config_default

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	69.372	66.111	16.624	0.048	3.645	7638.22	2366.6360320000003	72.4
256	35.475	33.175	8.485	0.047	3.603	7663.18	2366.6360320000003	72.6
128	18.478	16.646	3.96	0.045	3.544	7655.07	2366.6360320000003	72.4
64	9.279	8.315	1.584	0.044	3.448	7670.59	2366.6360320000003	72.0
32	7.395	4.977	0.617	0.041	3.28	6404.77	2366.6360320000003	71.6
16	5.849	3.927	0.699	0.04	2.93	4062.47	2383.413248	91.4
8	4.73	3.266	0.189	0.026	2.485	2439.19	2366.6360320000003	61.2
4	3.3	2.426	0.019	0.023	2.175	1640.9	2366.6360320000003	82.6
2	3.125	2.269	0.042	0.024	1.982	877.581	2299.5271679999996	48.7
1	3.108	2.619	0.055	0.027	2.271	379.415	2299.5271679999996	34.9

The model config "text_recognition_config_default" uses 4 GPU instances with a max batch size of 8 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.