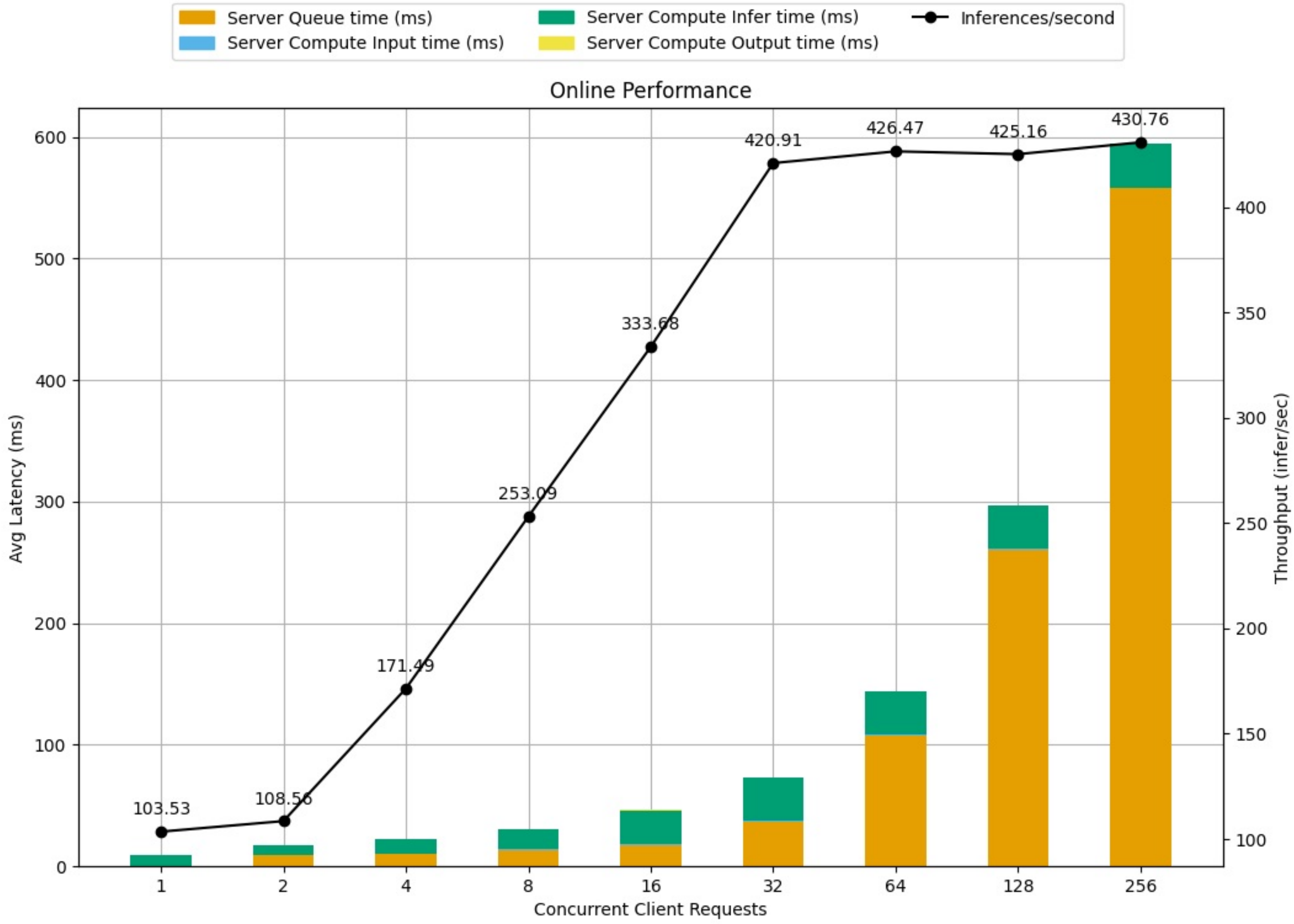
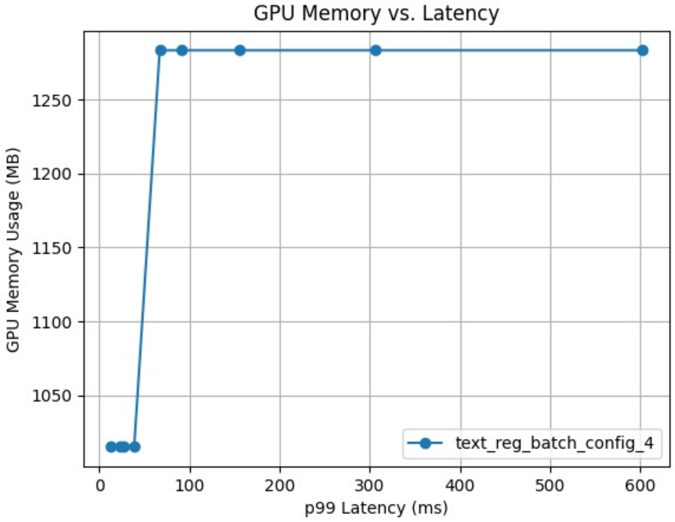
Detailed Report

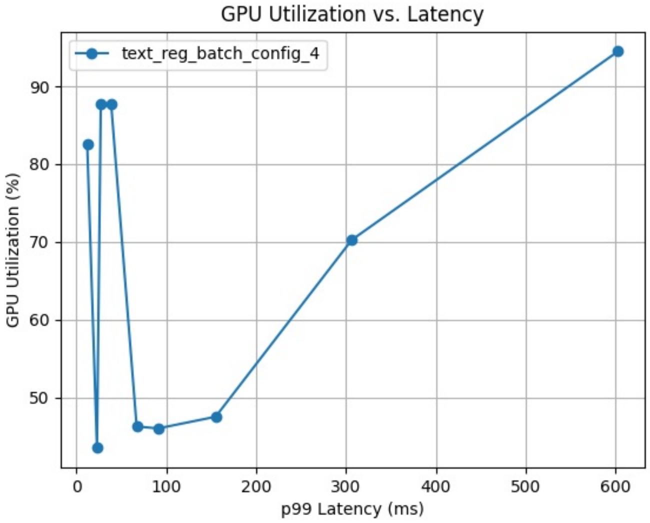
Model Config: text_reg_batch_config_4



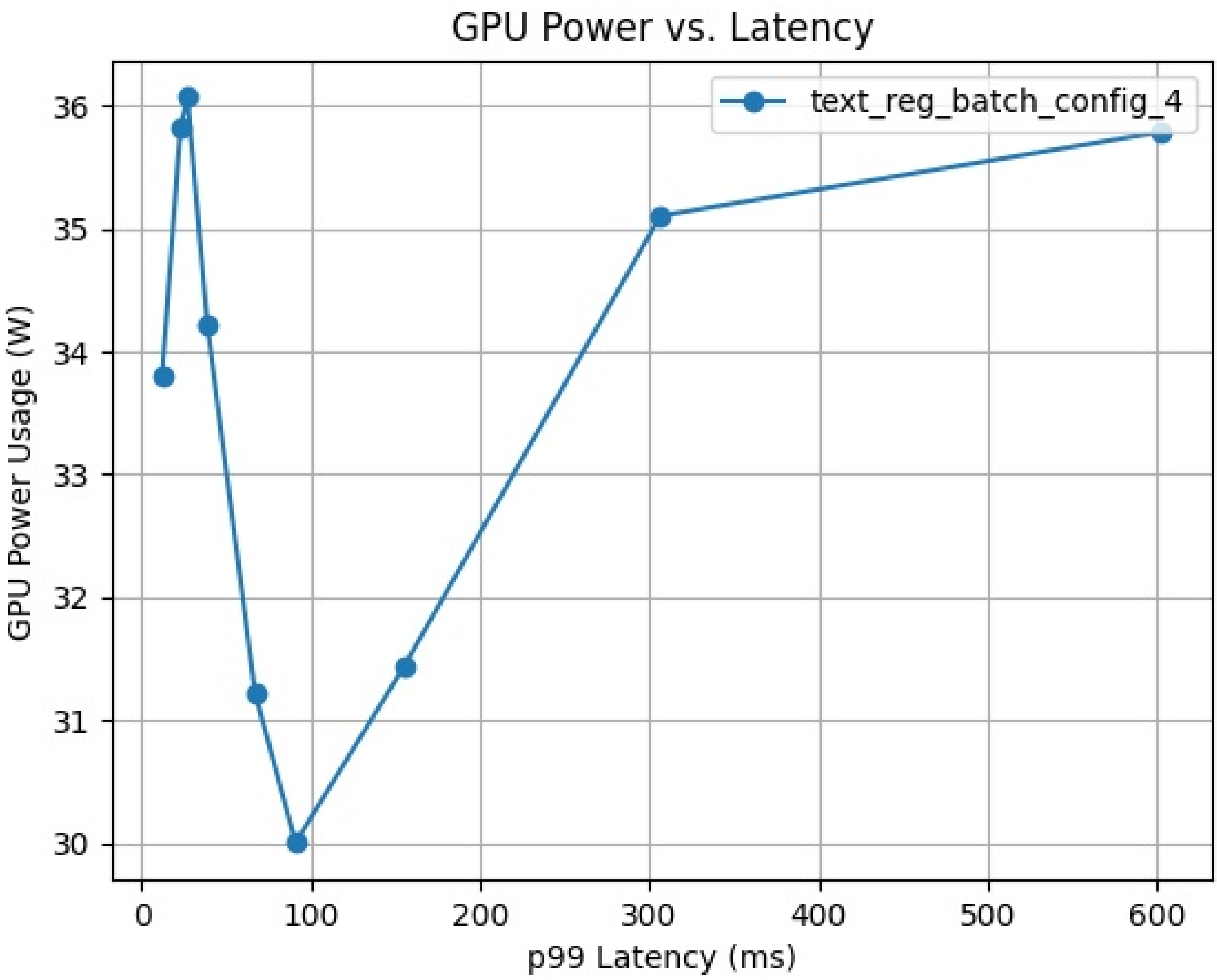
Latency Breakdown for Online Performance of text_reg_batch_config_4



GPU Memory vs. Latency curves for config text_reg_batch_config_4



GPU Utilization vs. Latency curves for config text_reg_batch_config_4



GPU Power vs. Latency curves for config text_reg_batch_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 (%)
~~256~~	603.053	595.67	558.135	0.312	35.708	430.757	1283.457024	94.5
128	306.568	298.356	260.542	0.313	35.866	425.16	1283.457024	70.2
64	154.918	145.356	107.712	0.378	35.613	426.473	1283.457024	47.5
32	90.885	75.085	36.785	0.378	35.998	420.911	1283.457024	46.0
16	67.01	47.466	17.718	0.276	28.022	333.676	1283.457024	46.2
8	38.67	31.6	13.57	0.197	16.884	253.087	1015.021568	87.7
4	27.078	23.218	9.983	0.142	12.292	171.486	1015.021568	87.7
2	22.833	18.283	8.684	0.056	8.903	108.556	1015.021568	43.5
1	11.981	9.608	0.142	0.062	8.854	103.534	1015.021568	82.5

The model config **text_reg_batch_config_4** uses 0 GPU instance with a max batch size of 16 and has dynamic batching enabled. 9 measurement(s) were obtained for the model config on GPU(s) with total memory 0 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 latency.