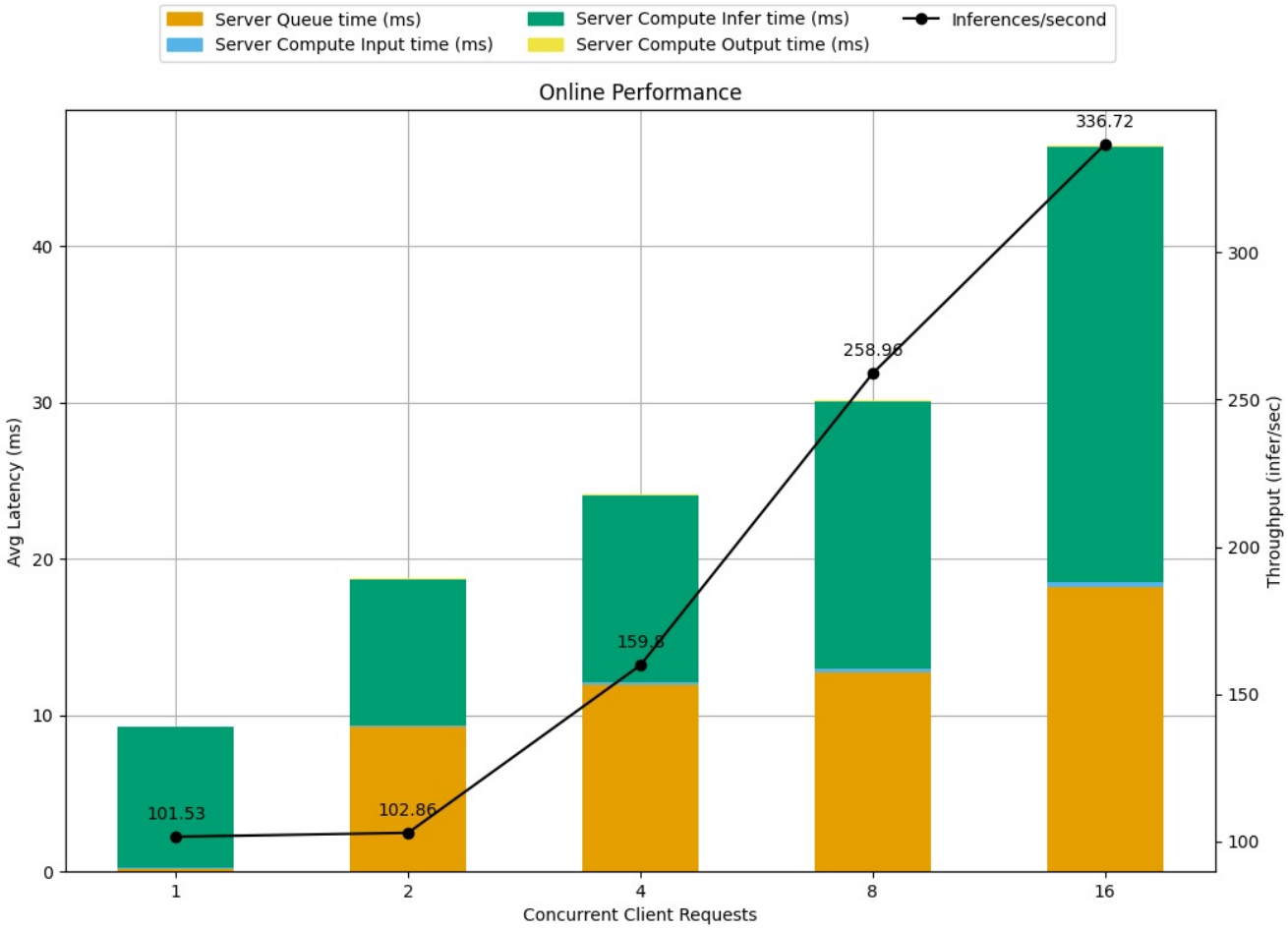
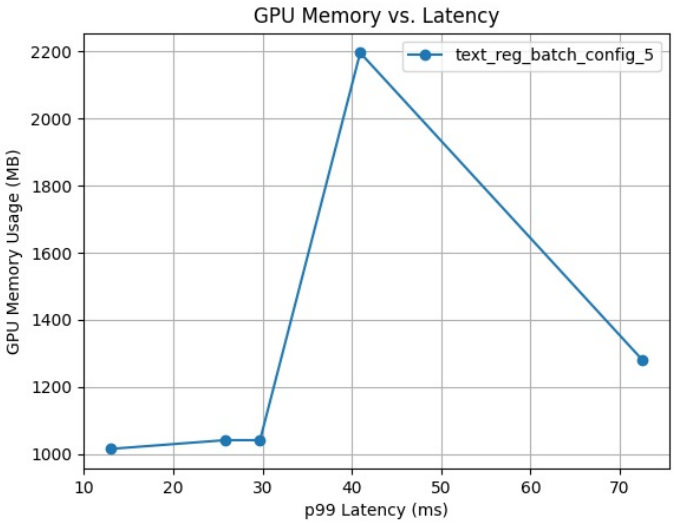
Detailed Report

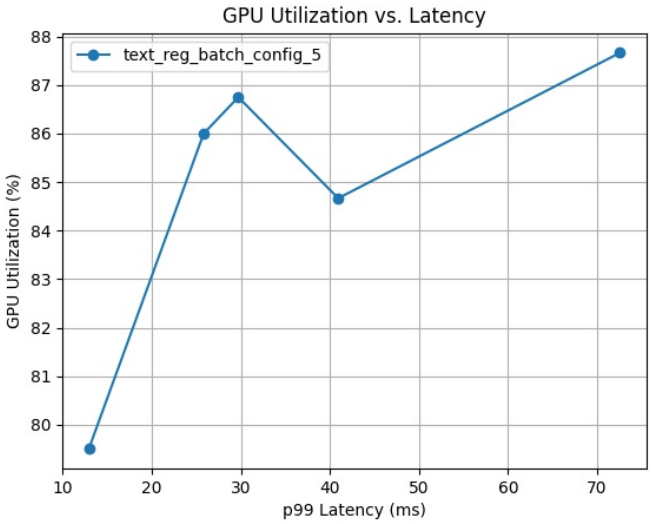
Model Config: text_reg_batch_config_5



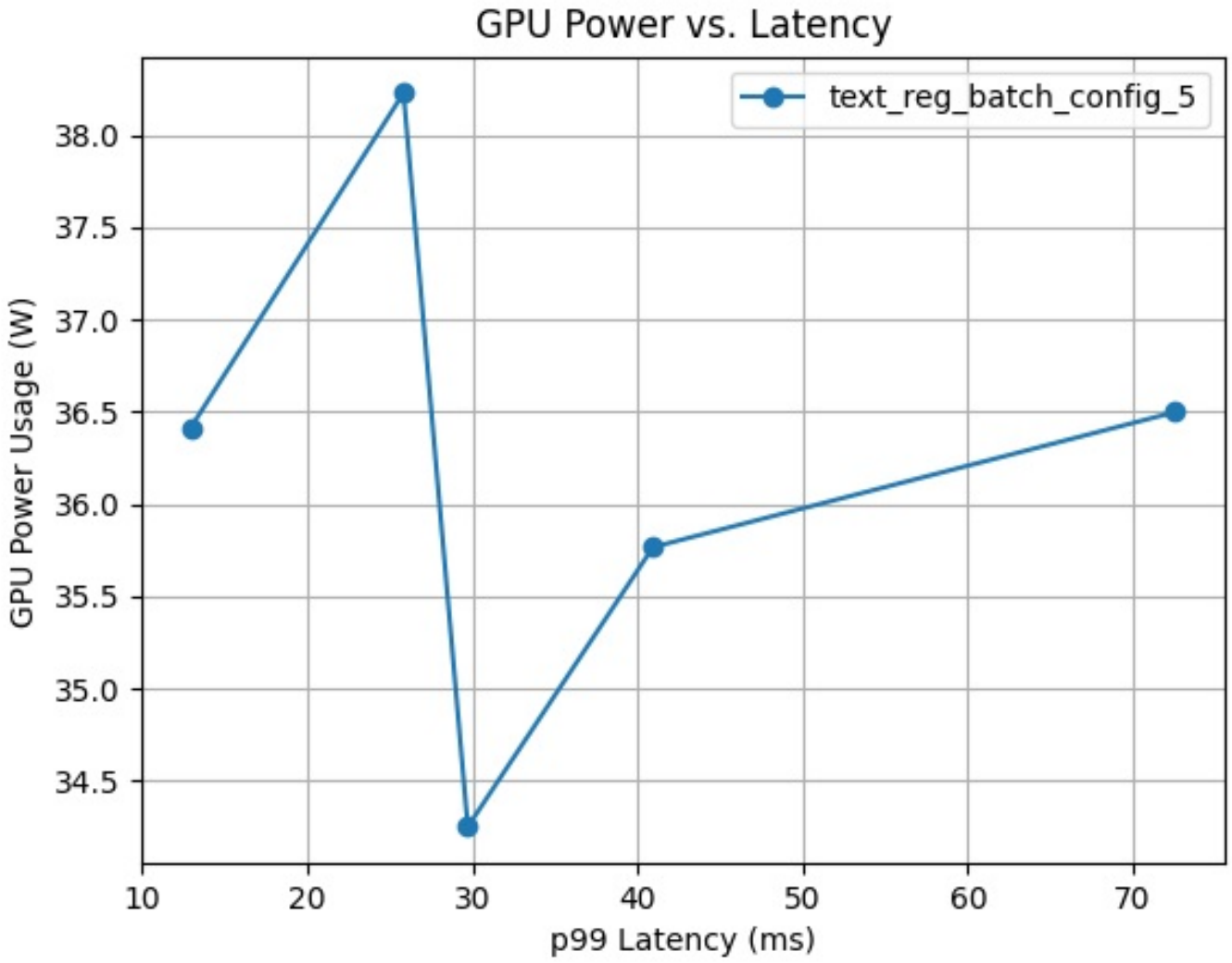
Latency Breakdown for Online Performance of text_reg_batch_config_5



GPU Memory vs. Latency curves for config text_reg_batch_config_5



GPU Utilization vs. Latency curves for config text_reg_batch_config_5



GPU Power vs. Latency curves for config text_reg_batch_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 (%)
16	72.569	47.56	18.219	0.283	27.841	336.716	1279.26272	87.7
8	40.926	30.943	12.775	0.188	17.096	258.961	2196.76672	84.7
4	29.709	24.888	11.918	0.176	11.979	159.801	1041.235968	86.8
2	25.815	19.378	9.25	0.063	9.369	102.864	1041.235968	86.0
1	12.92	9.781	0.164	0.06	9.015	101.531	1015.021568	79.5

The model config **text_reg_batch_config_5** uses 0 GPU instance with a max batch size of 32 and has dynamic batching enabled. 5 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.