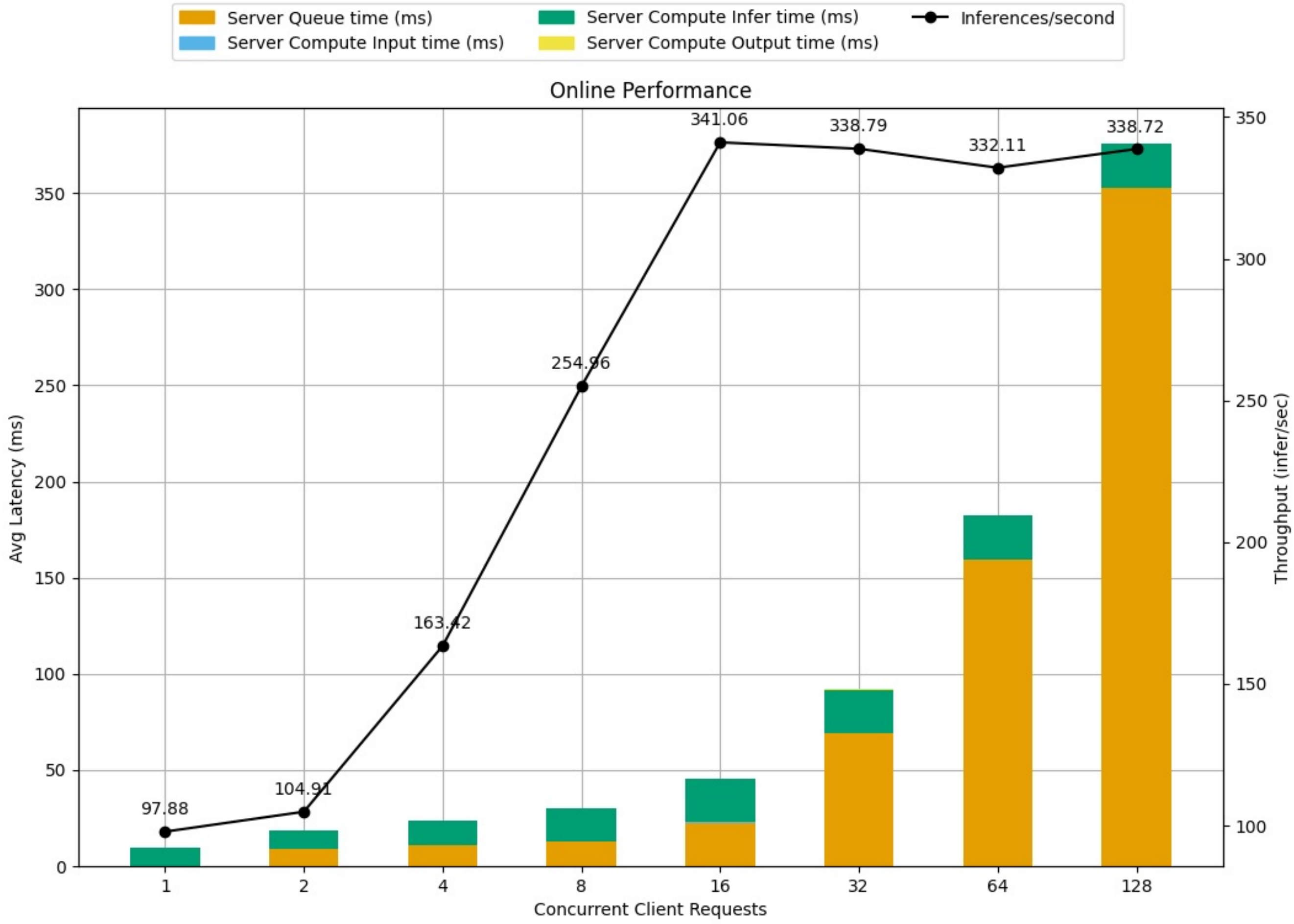
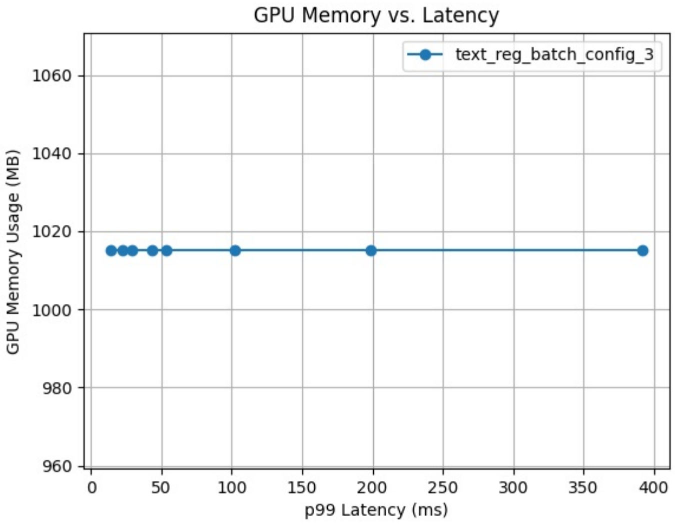
Detailed Report

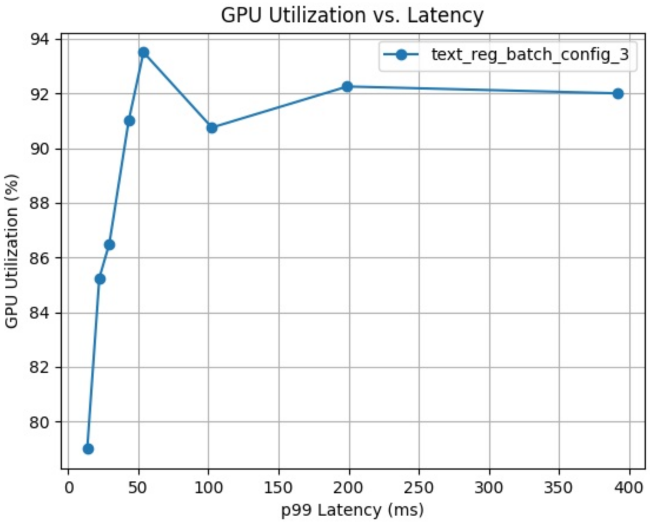
Model Config: text_reg_batch_config_3



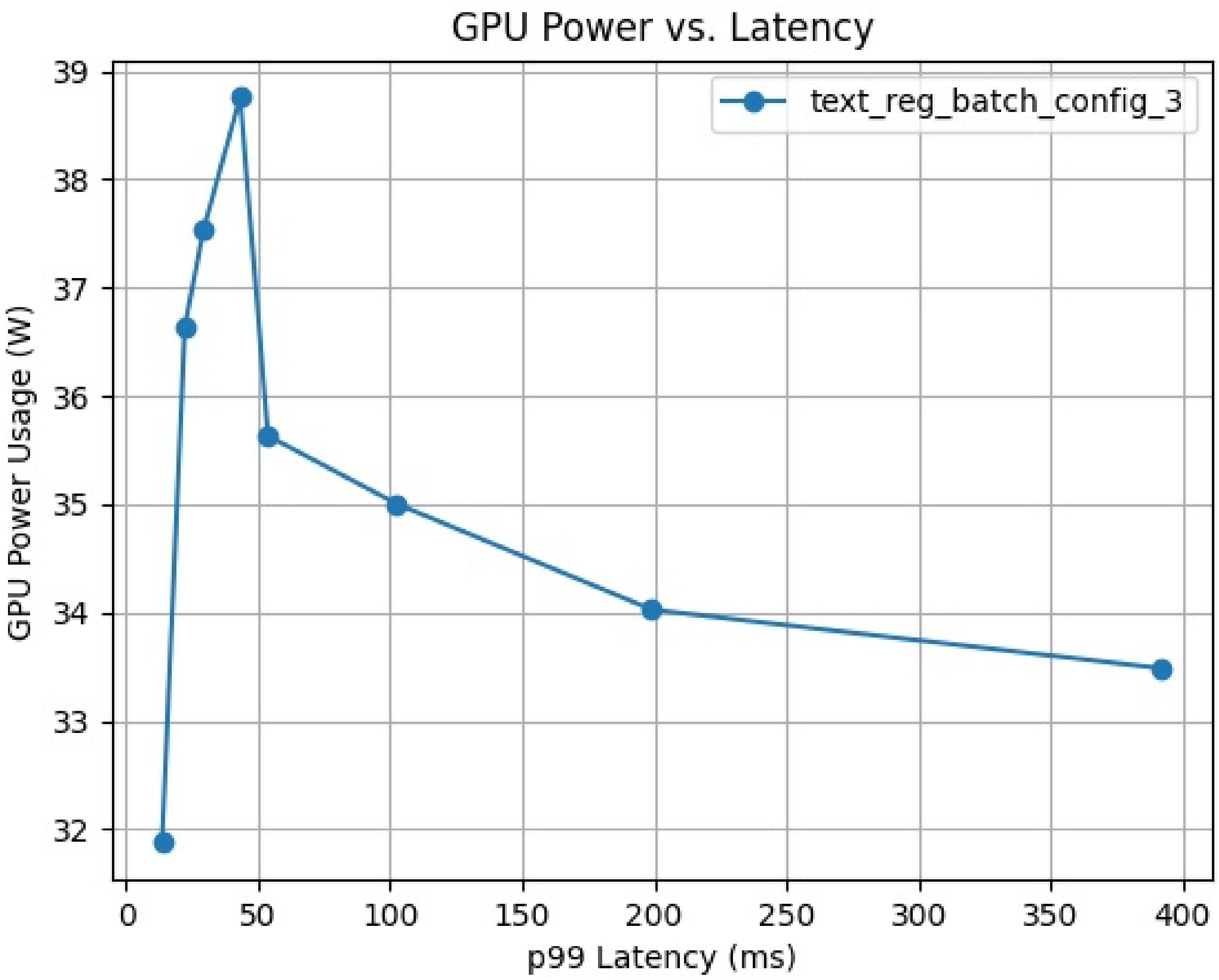
Latency Breakdown for Online Performance of text_reg_batch_config_3



GPU Memory vs. Latency curves for config text_reg_batch_config_3



GPU Utilization vs. Latency curves for config text_reg_batch_config_3



GPU Power vs. Latency curves for config text_reg_batch_config_3

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 (%)
128	392.082	376.724	352.713	0.25	22.463	338.717	1015.021568	92.0
64	199.036	183.527	159.312	0.276	22.548	332.106	1015.021568	92.2
32	102.514	93.013	68.997	0.379	22.432	338.793	1015.021568	90.8
16	53.717	46.432	22.709	0.298	22.284	341.062	1015.021568	93.5
8	43.181	31.256	12.885	0.205	17.256	254.964	1015.021568	91.0
4	29.362	24.342	10.979	0.145	12.373	163.422	1015.021568	86.5
2	22.252	18.942	9.012	0.066	9.208	104.905	1015.021568	85.2
1	13.617	10.181	0.153	0.073	9.327	97.8772	1015.021568	79.0

The model config **text_reg_batch_config_3** uses 0 GPU instance with a max batch size of 8 and has dynamic batching enabled. 8 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.