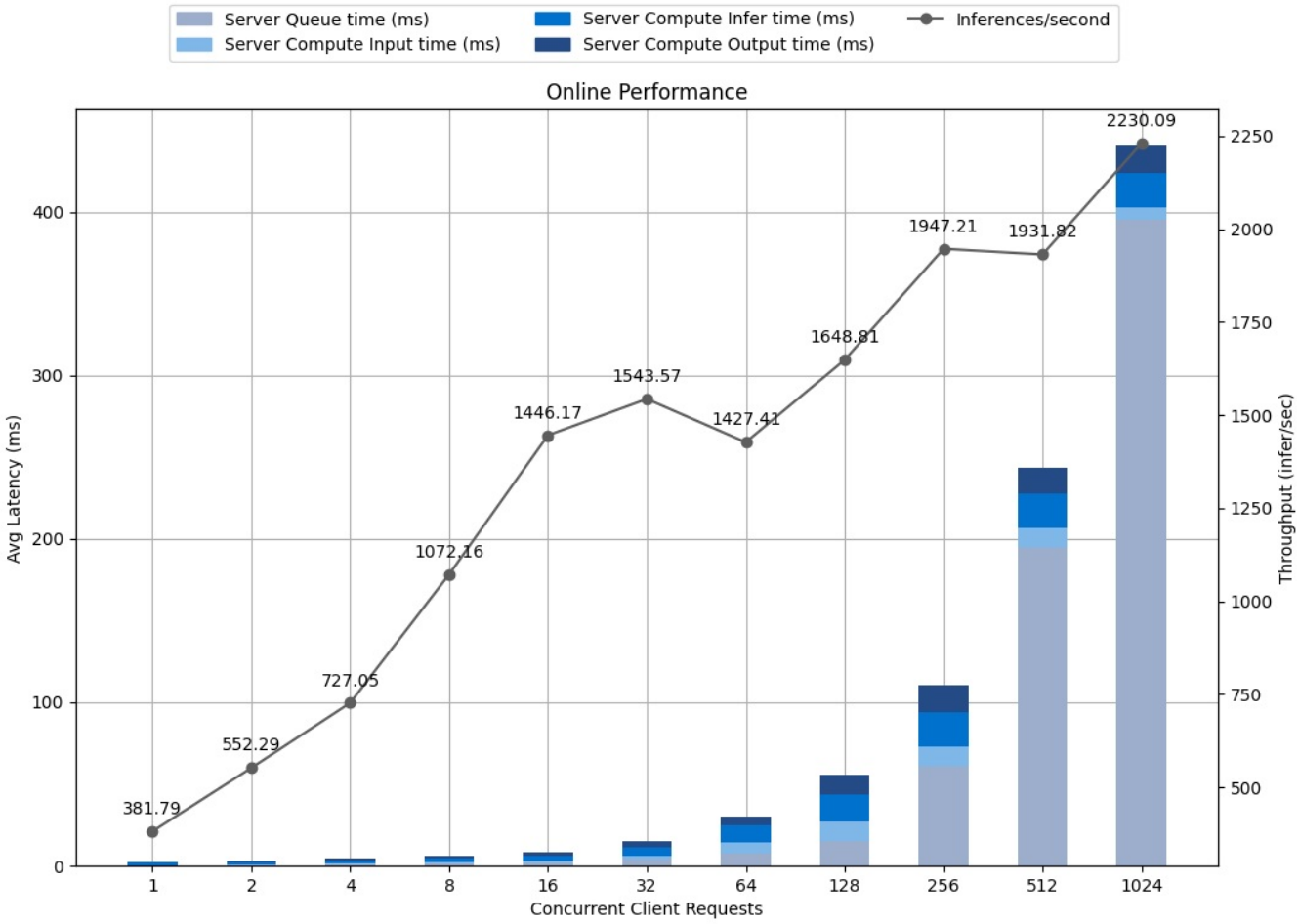
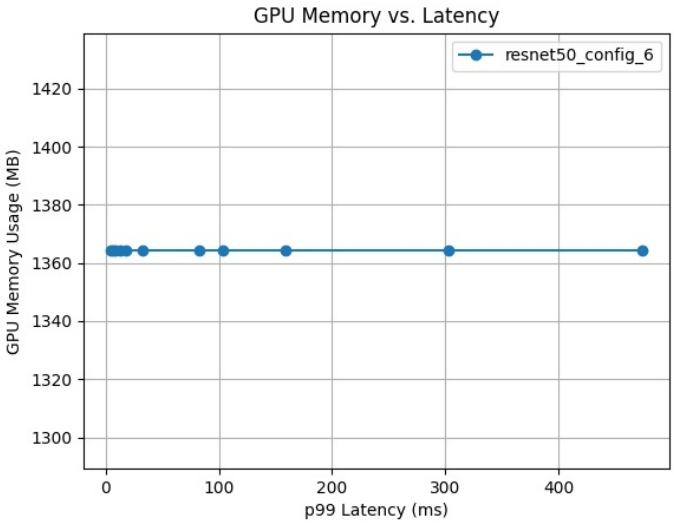
Detailed Report

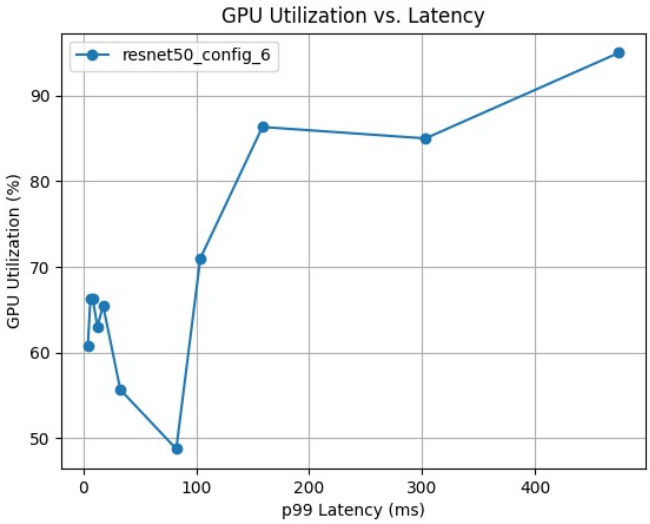
Model Config: resnet50_config_6



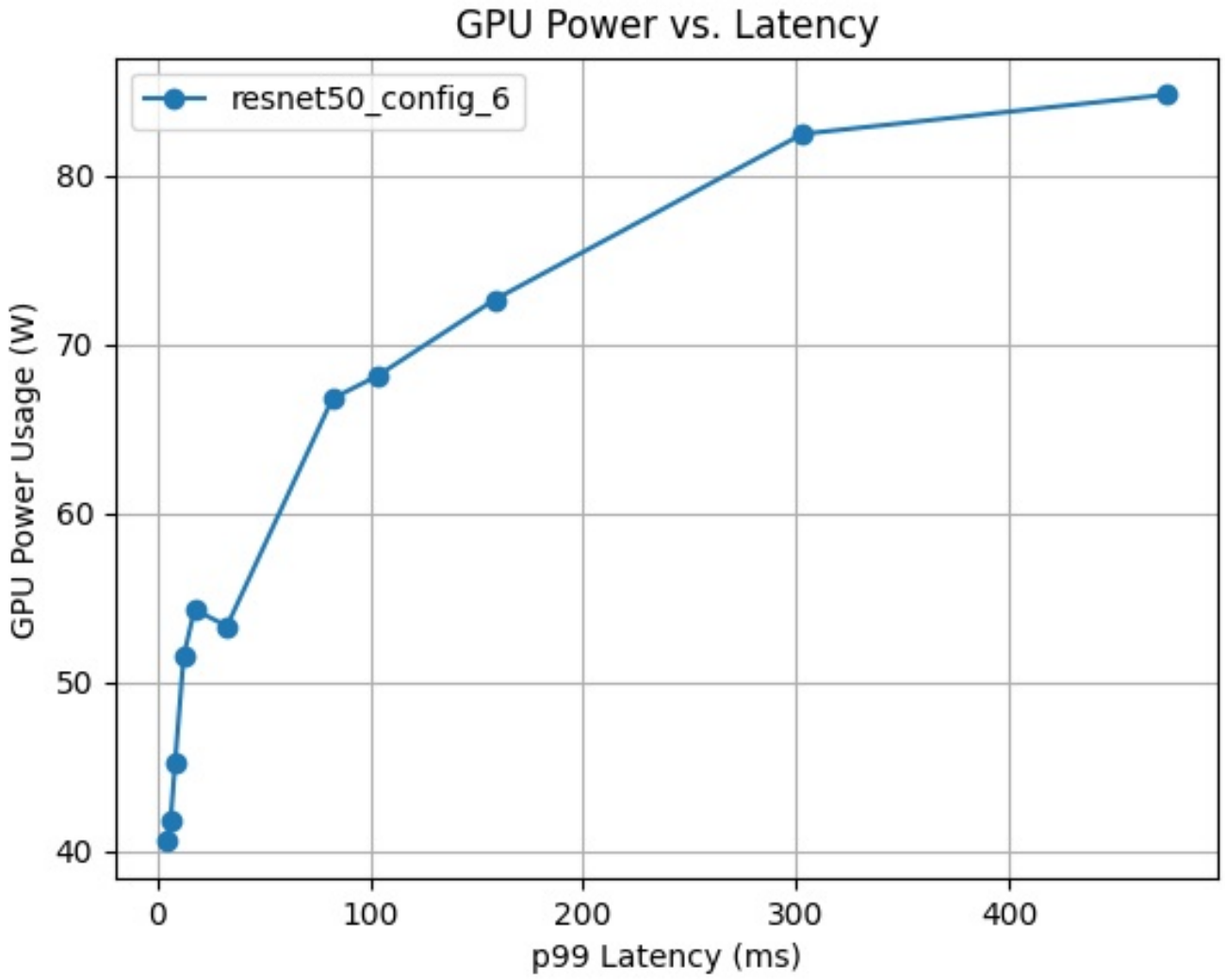
Latency Breakdown for Online Performance of resnet50_config_6



GPU Memory vs. Latency curves for config resnet50_config_6



GPU Utilization vs. Latency curves for config resnet50_config_6



GPU Power vs. Latency curves for config resnet50_config_6

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 (%)
1024	474.423	458.766	395.494	7.482	21.259	2230.09	1364.197376	95.0
512	302.994	263.969	194.779	12.107	21.078	1931.82	1364.197376	85.0
256	158.524	131.215	60.784	11.925	21.051	1947.21	1364.197376	86.3
128	103.563	76.91	15.117	11.581	17.199	1648.81	1364.197376	71.0
64	82.229	43.874	7.825	6.826	9.959	1427.41	1364.197376	48.8
32	32.684	20.35	3.404	2.52	5.542	1543.57	1364.197376	55.7
16	17.417	10.801	1.832	1.06	3.158	1446.17	1364.197376	65.5
8	12.364	7.301	1.491	0.609	2.316	1072.16	1364.197376	63.0
4	8.354	5.392	1.01	0.259	1.761	727.055	1364.197376	66.2
2	6.21	3.52	0.222	0.173	1.608	552.295	1364.197376	66.2
1	3.877	2.552	0.086	0.259	1.605	381.788	1364.197376	60.8

The model config **resnet50_config_6** uses 1 GPU instance with a max batch size of 64 and has dynamic batching enabled. 11 measurement(s) were obtained for the model config on GPU(s) 1 x NVIDIA GeForce RTX 3060 Laptop GPU with total memory 6.0 GB. This model uses the platform tensorrt_plan.

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.