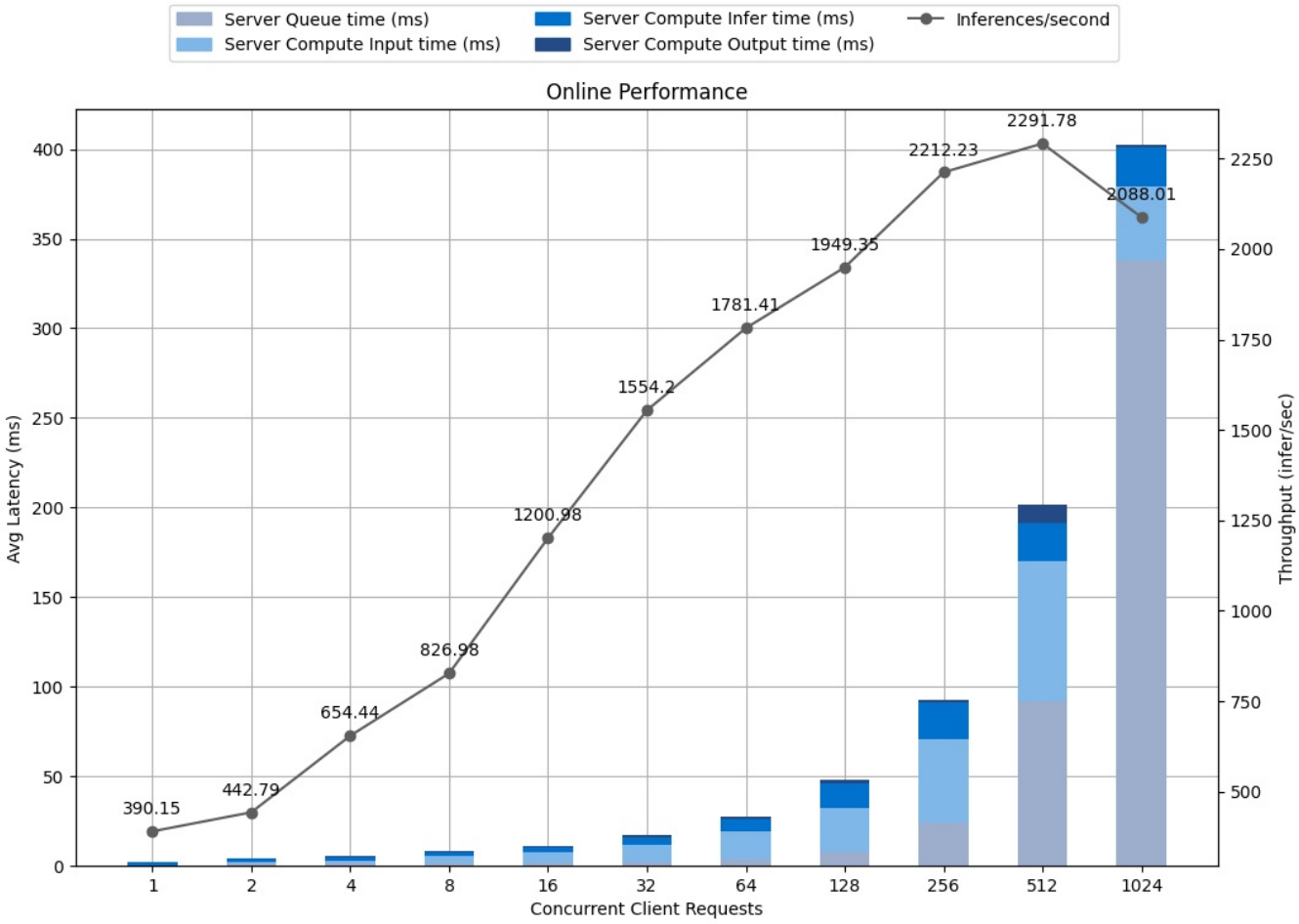
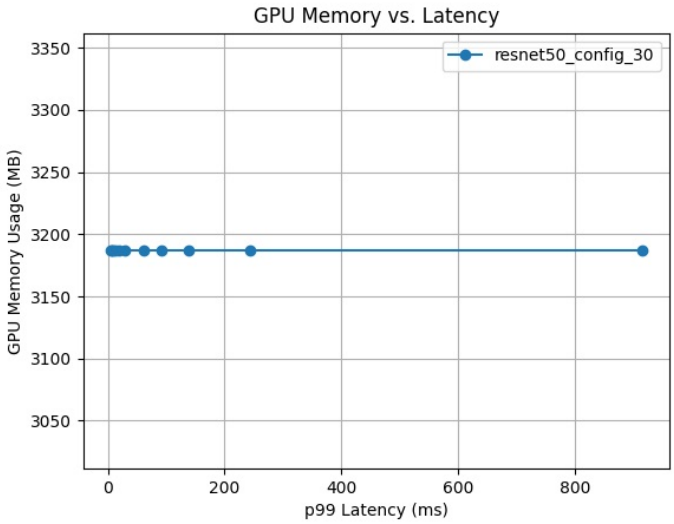
Detailed Report

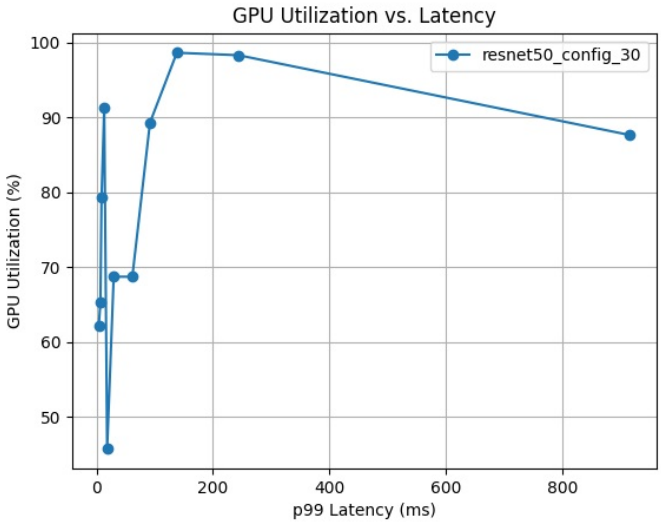
Model Config: resnet50_config_30



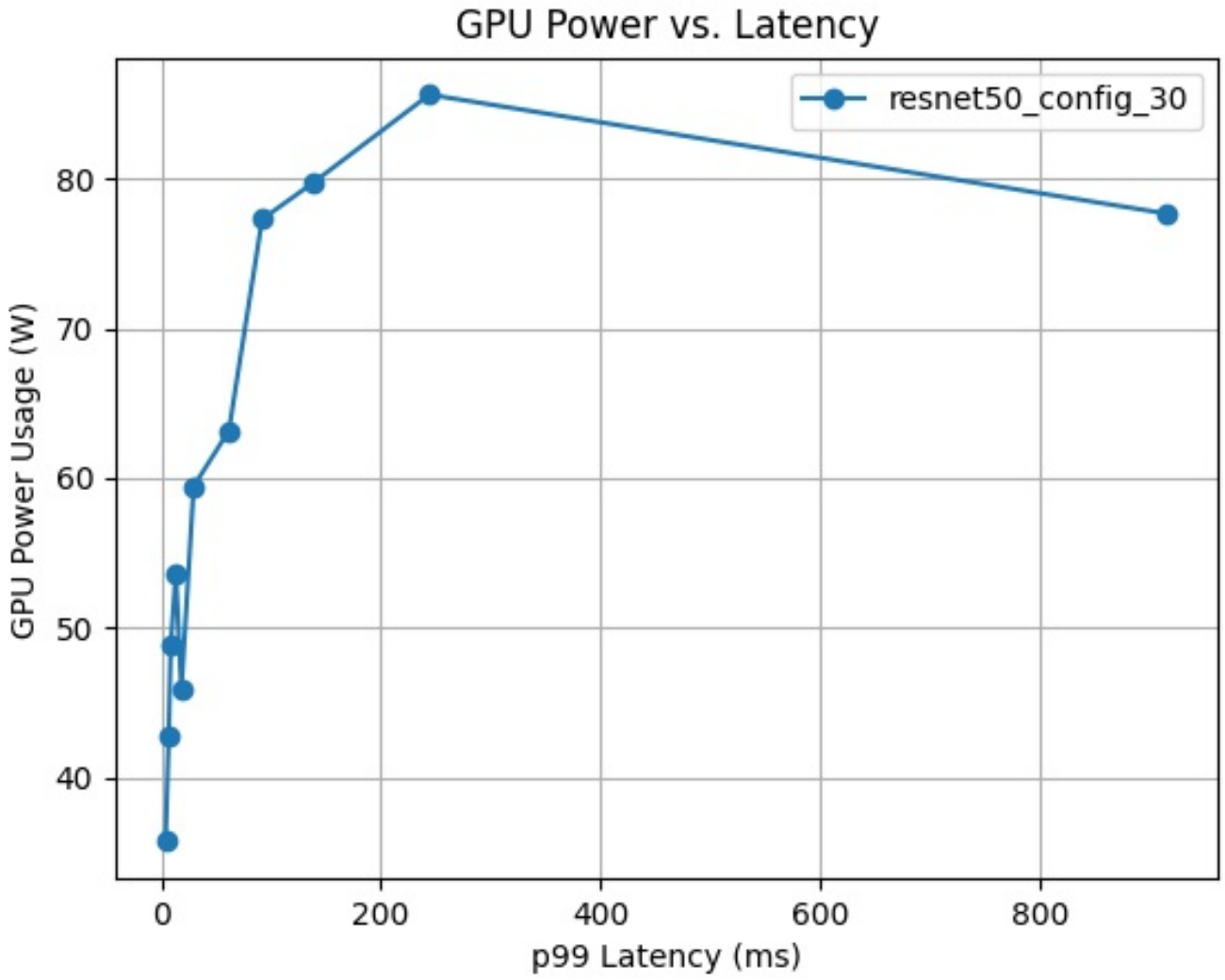
Latency Breakdown for Online Performance of resnet50_config_30



GPU Memory vs. Latency curves for config resnet50_config_30



GPU Utilization vs. Latency curves for config resnet50_config_30



GPU Power vs. Latency curves for config resnet50_config_30

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	916.502	493.487	338.083	41.062	22.115	2088.01	3186.622464	87.7
512	244.039	223.086	92.123	77.72	21.242	2291.78	3186.622464	98.3
256	137.81	115.483	23.729	46.928	20.295	2212.23	3186.622464	98.7
128	91.68	65.245	7.662	24.808	13.618	1949.35	3186.622464	89.3
64	61.366	35.108	3.431	15.644	6.672	1781.41	3186.622464	68.8
32	29.333	20.3	1.898	9.977	3.556	1554.2	3186.622464	68.8
16	18.054	13.138	1.362	6.355	2.477	1200.98	3186.622464	45.8
8	13.054	9.583	1.101	4.576	1.852	826.981	3186.622464	91.3
4	8.429	6.034	1.063	1.917	1.911	654.438	3186.622464	79.3
2	6.536	4.457	1.48	0.38	1.954	442.786	3186.622464	65.3
1	3.743	2.503	0.086	0.235	1.593	390.152	3186.622464	62.3

The model config **resnet50_config_30** uses 4 GPU instances 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.