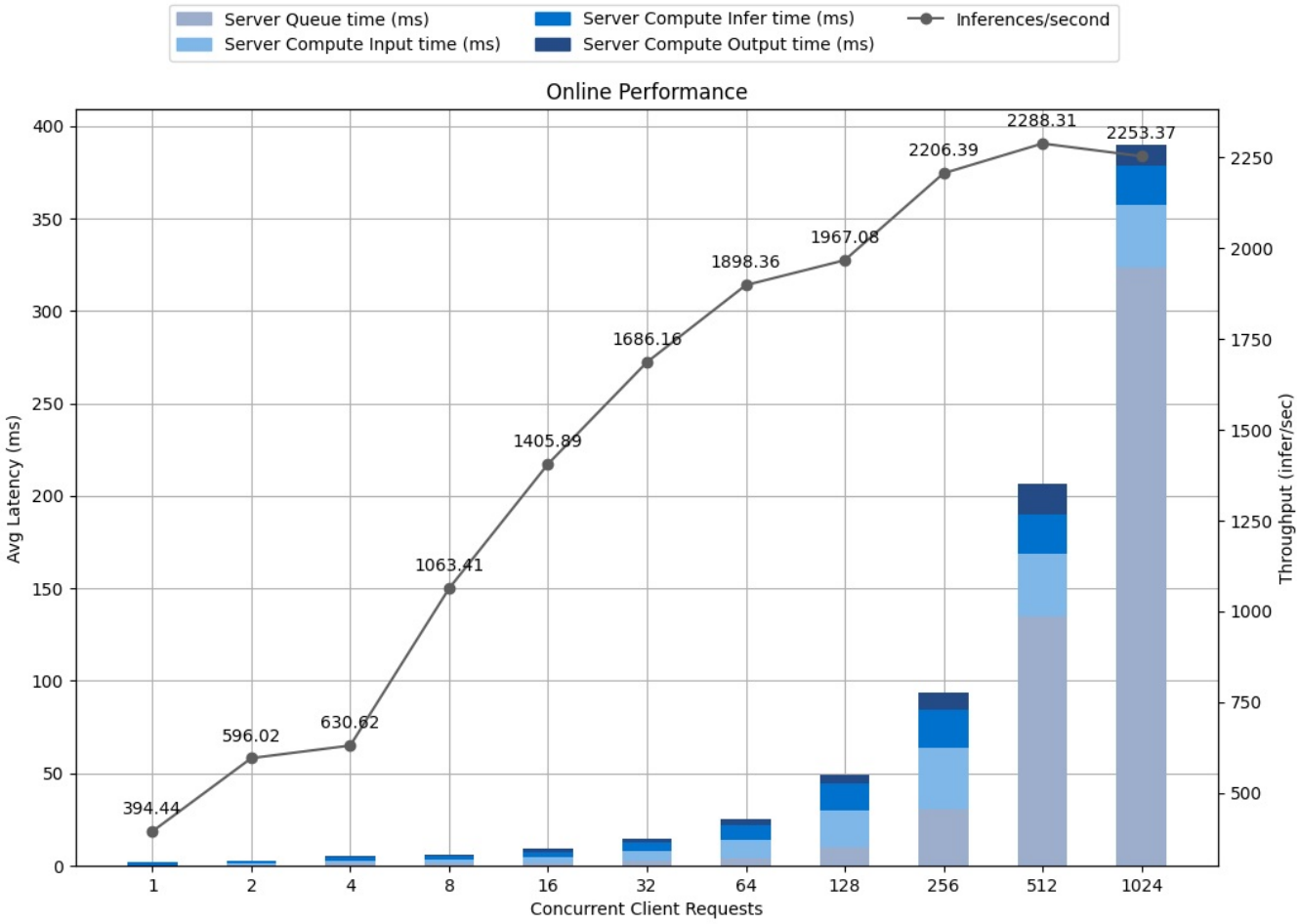
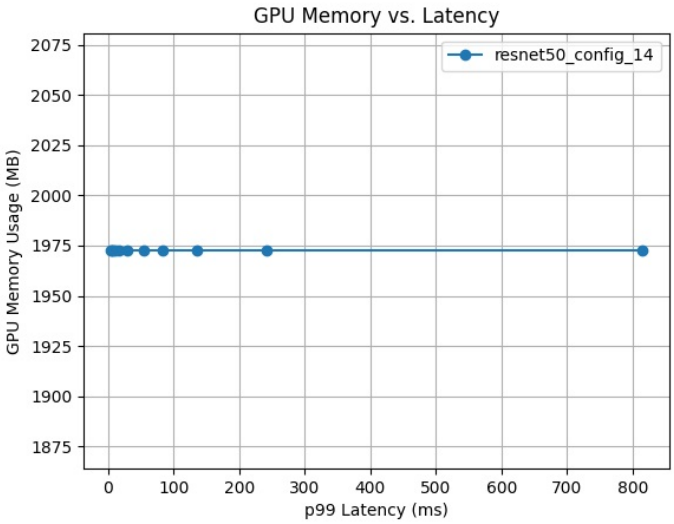
Detailed Report

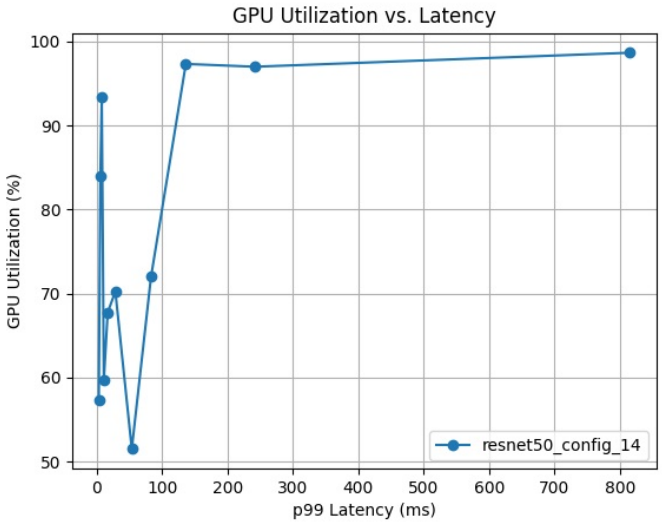
Model Config: resnet50_config_14



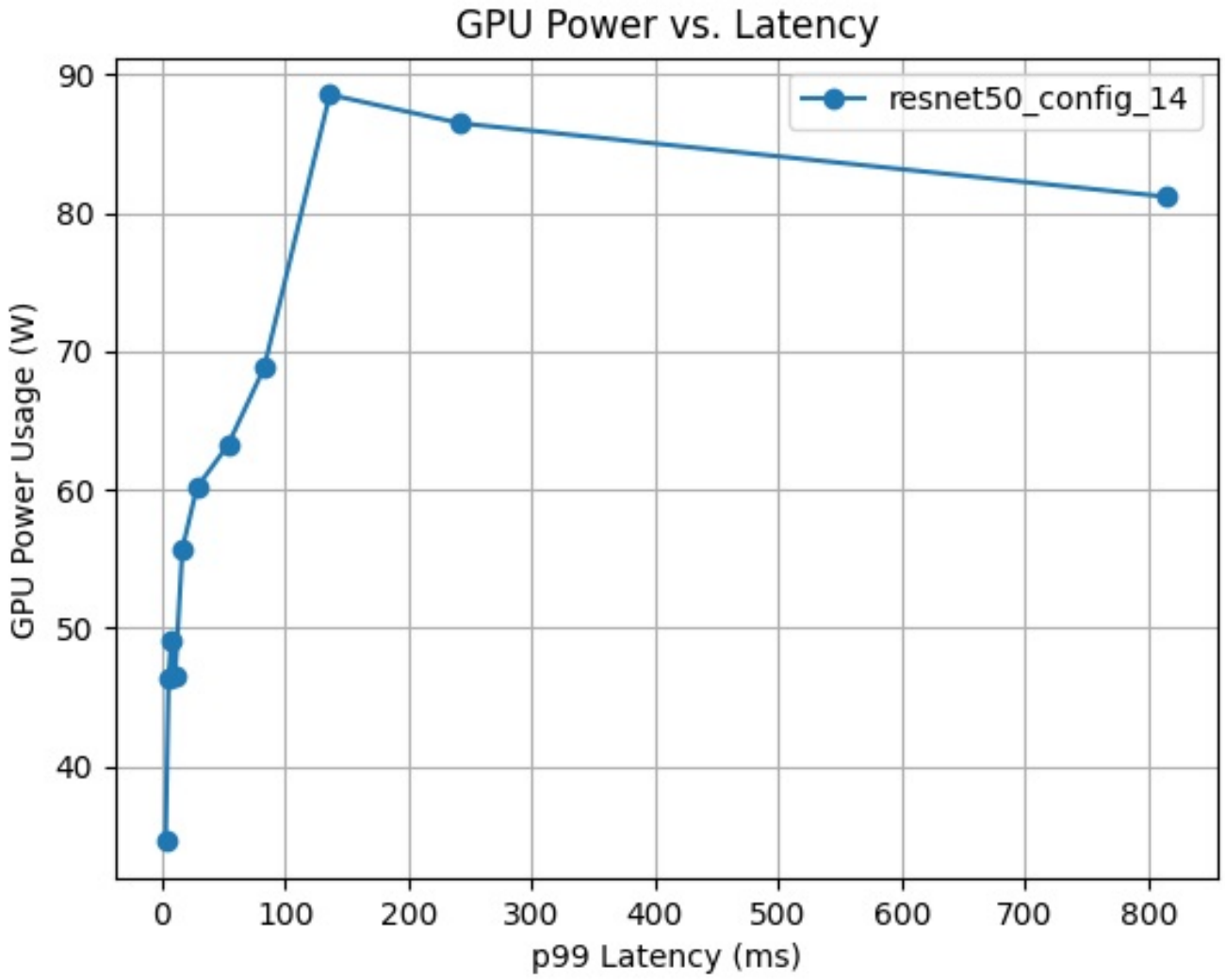
Latency Breakdown for Online Performance of resnet50_config_14



GPU Memory vs. Latency curves for config resnet50_config_14



GPU Utilization vs. Latency curves for config resnet50_config_14



GPU Power vs. Latency curves for config resnet50_config_14

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	815.471	460.141	323.536	33.575	21.265	2253.37	1972.371456	98.7
512	242.088	224.318	134.697	34.165	21.17	2288.31	1972.371456	97.0
256	136.307	115.505	30.4	33.078	21.126	2206.39	1972.371456	97.3
128	83.092	64.957	9.689	20.093	14.782	1967.08	1972.371456	72.0
64	53.662	33.0	4.227	9.783	7.782	1898.36	1972.371456	51.5
32	28.556	18.678	2.486	5.407	4.56	1686.16	1972.371456	70.2
16	16.856	11.199	1.592	3.158	2.775	1405.89	1972.371456	67.8
8	10.917	7.386	1.456	1.609	2.347	1063.41	1972.371456	59.7
4	7.783	6.282	1.129	1.674	1.555	630.625	1972.371456	93.3
2	5.638	3.299	0.193	0.999	1.515	596.024	1972.371456	84.0
1	3.237	2.484	0.088	0.232	1.554	394.437	1972.371456	57.2

The model config **resnet50_config_14** uses 2 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.