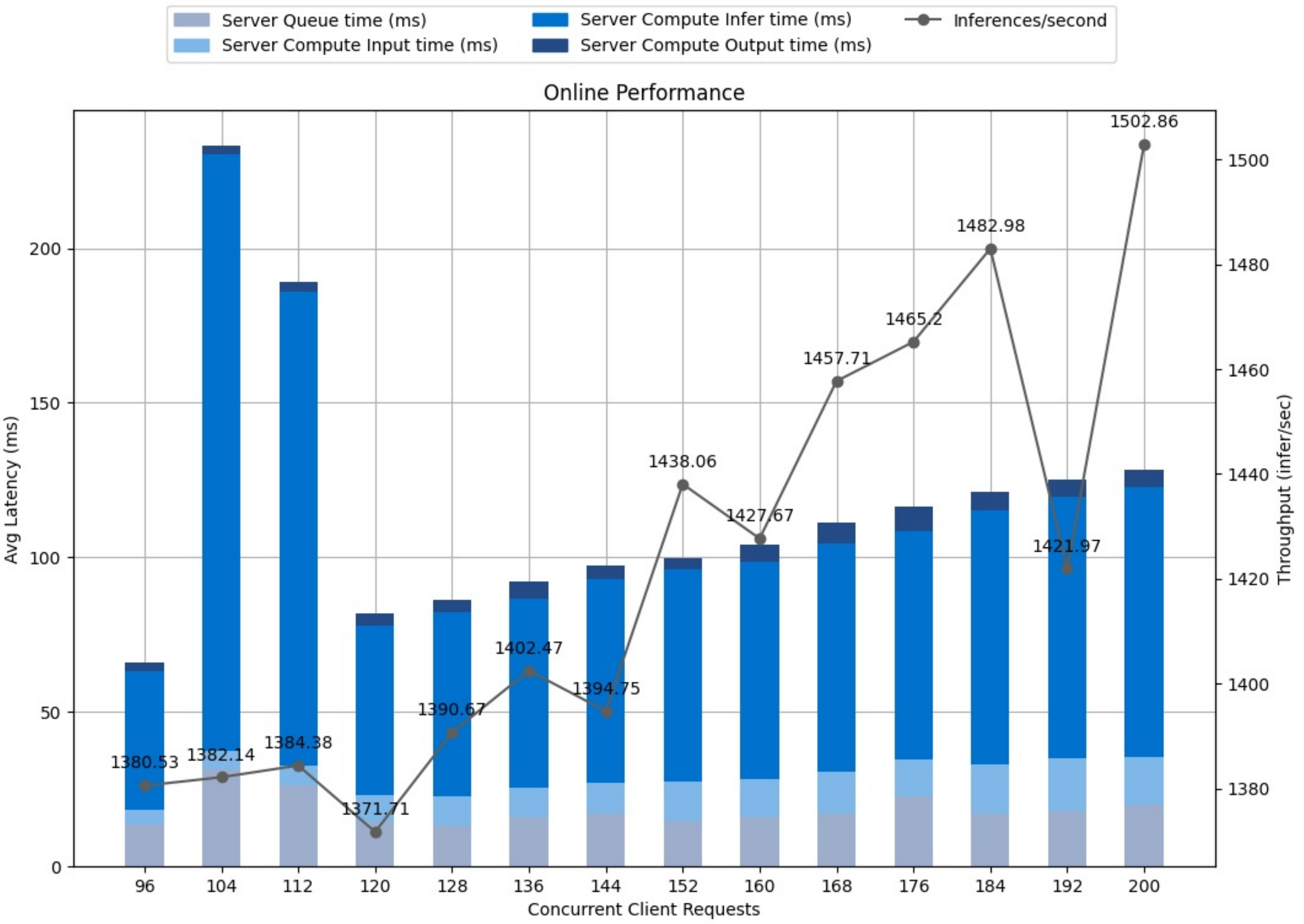
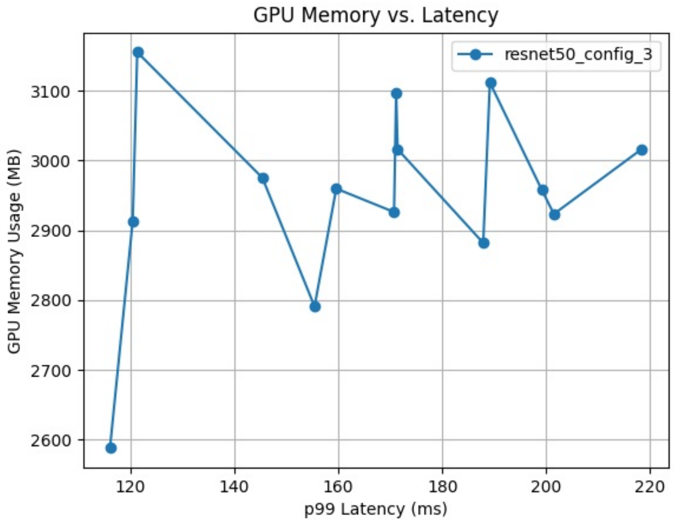
Detailed Report

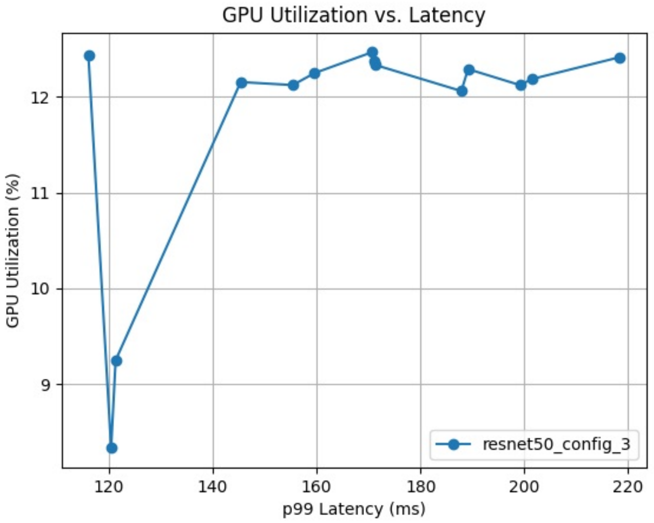
Model Config: resnet50_config_3



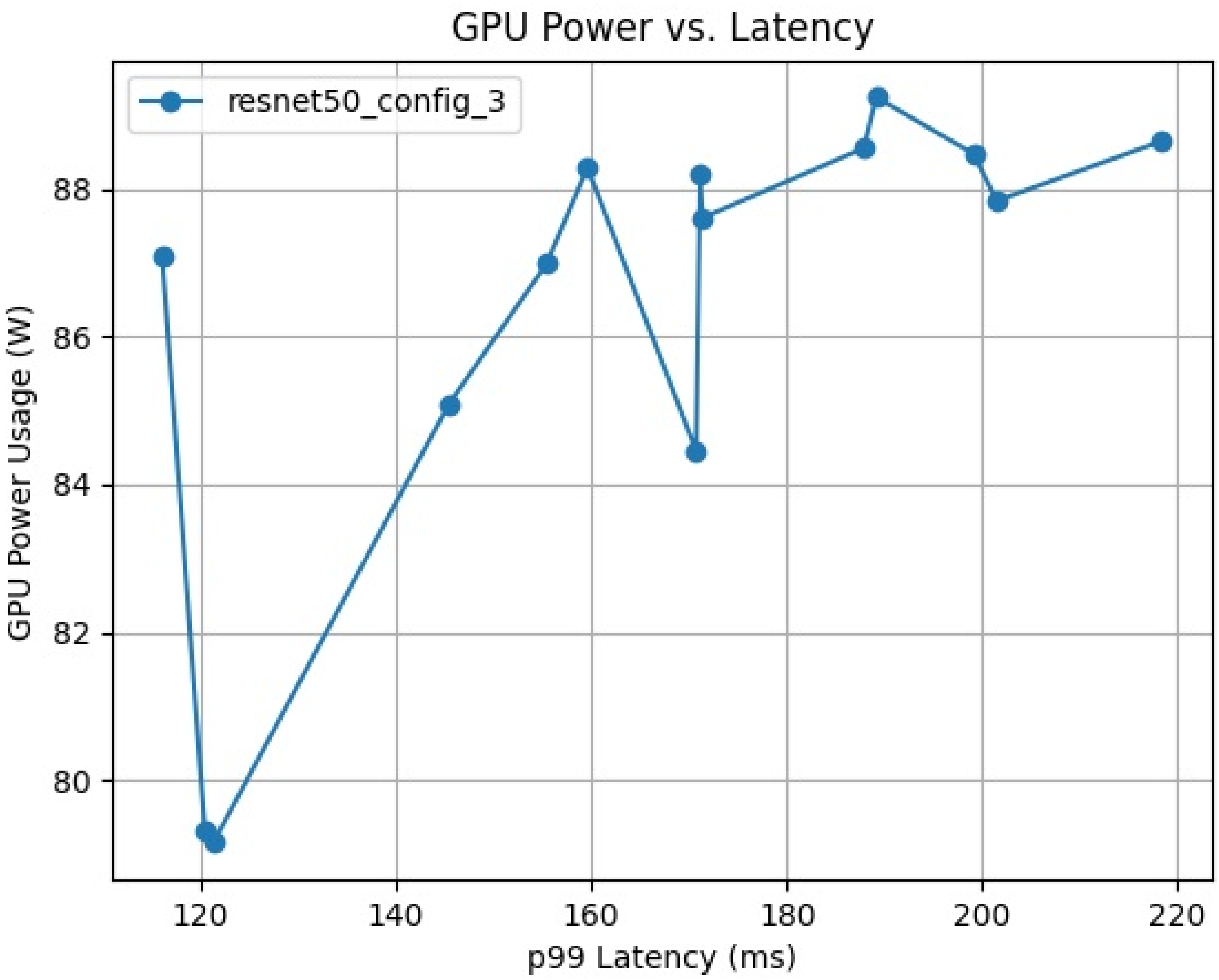
Latency Breakdown for Online Performance of resnet50_config_3



GPU Memory vs. Latency curves for config resnet50_config_3



GPU Utilization vs. Latency curves for config resnet50_config_3



GPU Power vs. Latency curves for config resnet50_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 (%)
192	218.414	130.969	17.861	17.246	84.424	1421.97	3016.622080000001	12.4
168	201.47	117.491	17.07	13.64	73.881	1457.71	2923.298816000001	12.2
200	199.152	135.76	19.781	15.71	87.412	1502.86	2957.901824000001	12.1
184	189.16	127.504	17.276	15.657	82.168	1482.98	3111.780352000001	12.3
176	187.879	121.189	22.797	11.966	73.752	1465.2	2881.880064000001	12.1
160	171.41	109.637	15.899	12.468	70.011	1427.67	3016.097792000001	12.3
144	171.124	101.932	17.001	9.934	66.055	1394.75	3096.313856000001	12.4
152	170.722	105.208	14.896	12.388	68.672	1438.06	2926.1824000000006	12.5
120	159.601	85.593	14.377	8.541	54.8	1371.71	2959.736832000001	12.2
136	155.423	96.572	15.934	9.496	61.263	1402.47	2790.9160960000004	12.1
128	145.39	90.962	13.047	9.688	59.459	1390.67	2975.203328000001	12.2
112	121.311	79.232	26.201	6.354	153.26	1384.38	3155.296256000001	9.2
104	120.442	73.763	31.285	5.969	193.005	1382.14	2913.5994880000007	8.3
96	116.064	68.896	13.497	4.963	44.593	1380.53	2588.0166400000003	12.4

The model config "resnet50_config_3" uses 8 GPU instances with a max batch size of 128 and has dynamic batching enabled. 14 measurement(s) were obtained for the model config on GPU(s) 8 x Tesla V100-SXM2-16GB with total memory 126.4 GB. This model uses the platform pytorch_libtorch.

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 throughput.