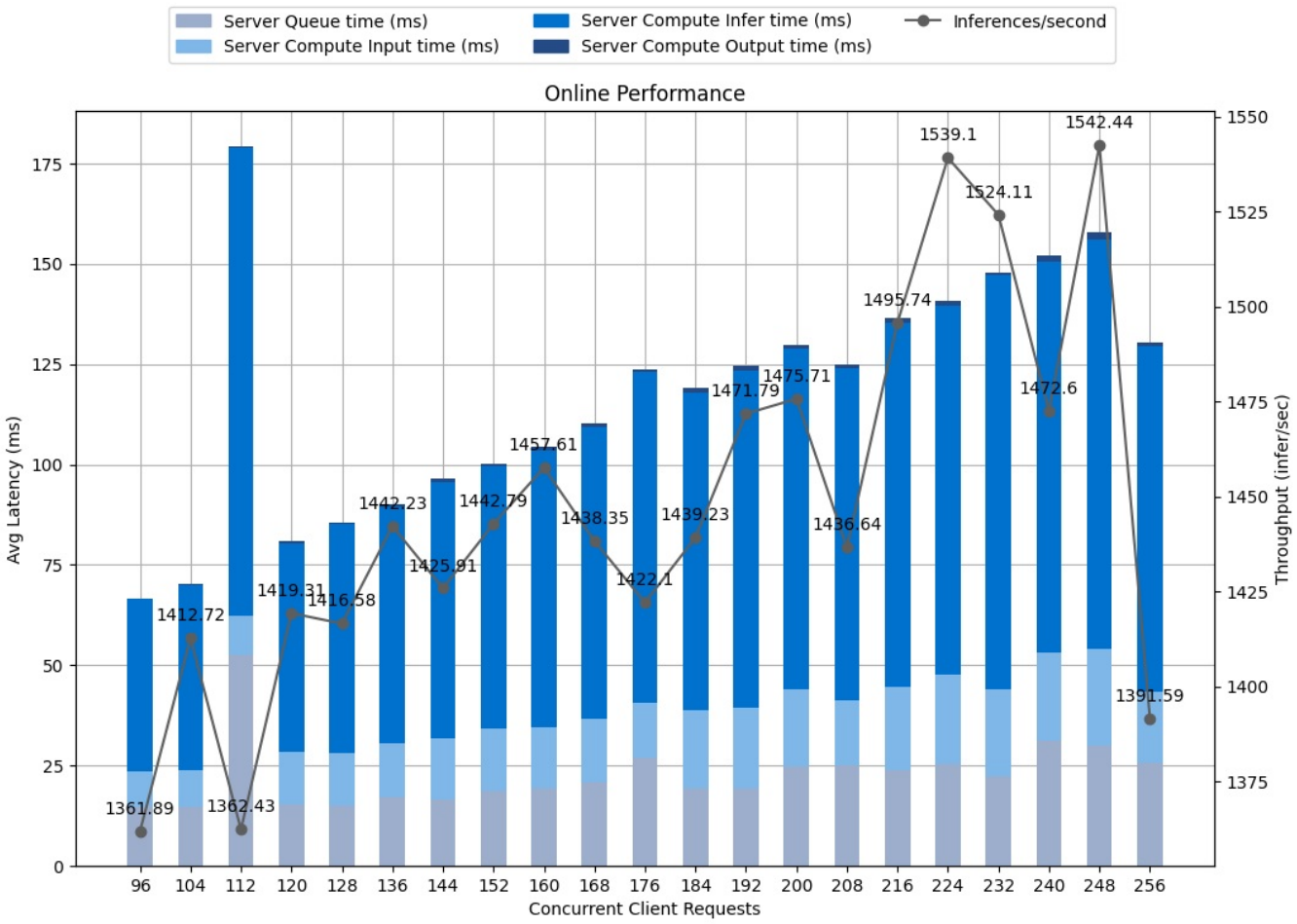
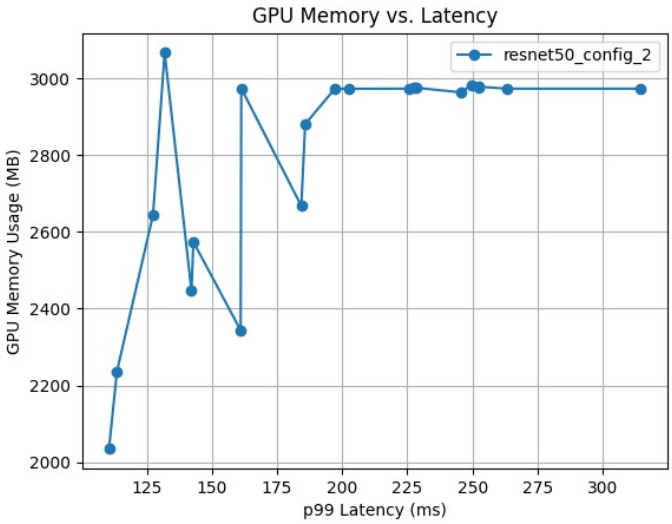
Detailed Report

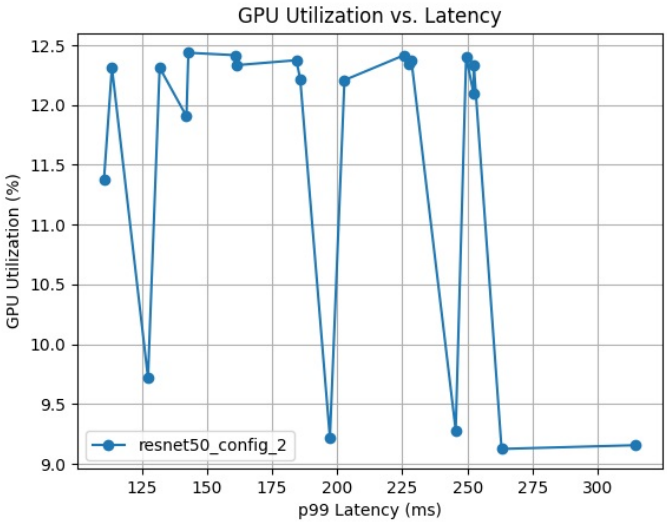
Model Config: resnet50_config_2



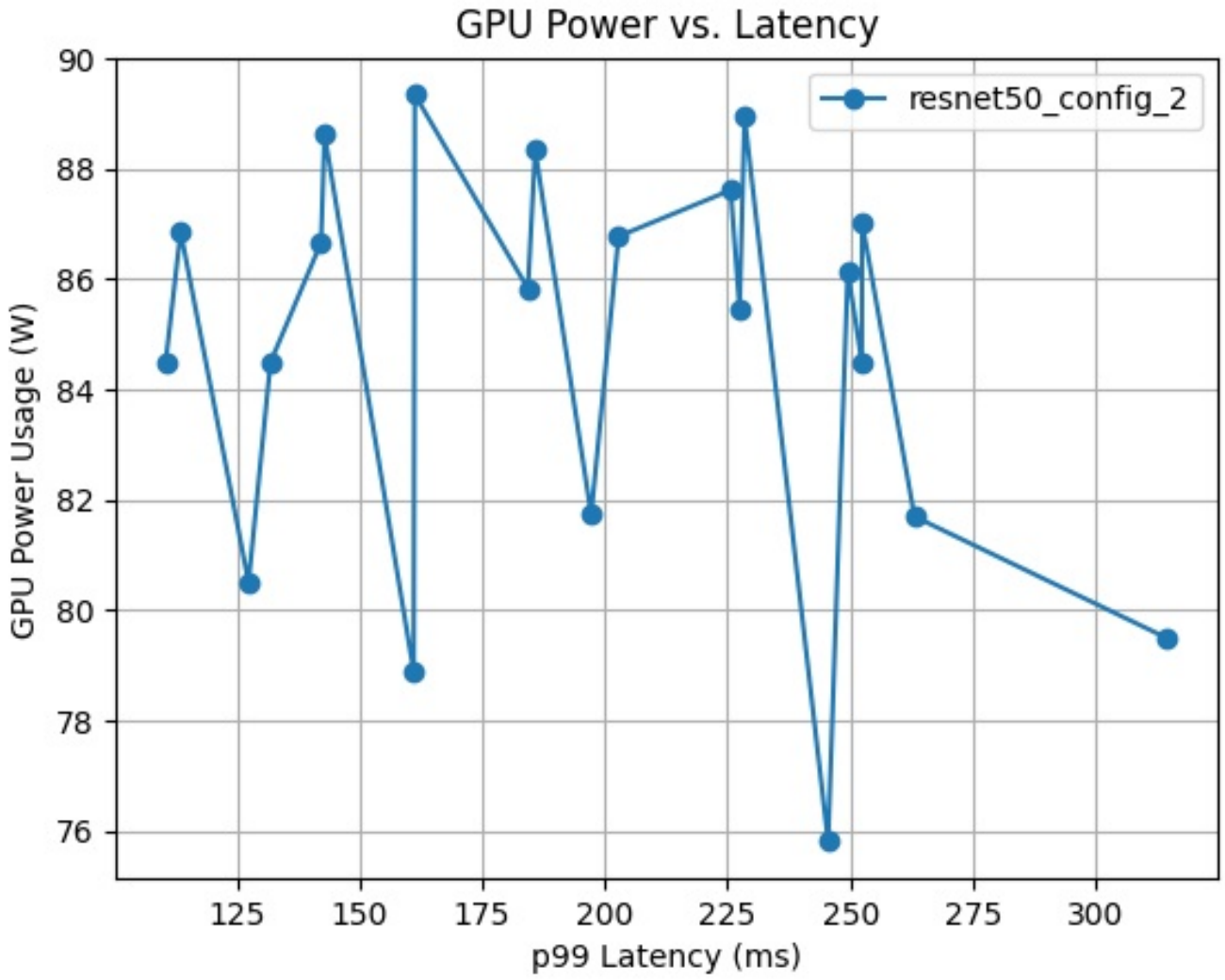
Latency Breakdown for Online Performance of resnet50_config_2



GPU Memory vs. Latency curves for config resnet50_config_2



GPU Utilization vs. Latency curves for config resnet50_config_2



GPU Power vs. Latency curves for config resnet50_config_2

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 (%)
256	314.628	170.317	25.628	17.75	86.163	1391.59	2972.975104000001	9.2
208	263.088	139.231	25.149	16.152	82.841	1436.64	2972.975104000001	9.1
232	252.436	157.718	22.169	21.842	103.152	1524.11	2978.217984000001	12.3
240	252.355	161.478	31.243	21.891	97.493	1472.6	2977.169408000001	12.1
248	249.504	167.74	30.027	24.043	102.016	1542.44	2981.3637120000008	12.4
176	245.493	119.981	26.962	13.552	82.666	1422.1	2963.275776000001	9.3
200	228.475	137.432	24.781	19.341	84.653	1475.71	2975.334400000001	12.4
216	227.468	145.394	23.813	20.748	90.82	1495.74	2975.596544000001	12.3
224	225.654	148.712	25.448	22.278	91.748	1539.1	2972.975104000001	12.4
184	202.808	126.719	19.202	19.594	79.232	1439.23	2972.7129600000007	12.2
192	197.23	132.231	19.366	20.02	83.869	1471.79	2972.975104000001	9.2
160	185.822	110.806	19.264	15.115	69.243	1457.61	2880.7004160000006	12.2
168	184.336	117.19	20.677	16.062	72.486	1438.35	2667.3152000000005	12.4
144	161.301	102.45	16.56	15.169	63.843	1425.91	2973.761536000001	12.3
152	161.071	106.159	18.583	15.516	65.342	1442.79	2343.3052159999997	12.4
136	142.915	95.32	17.101	13.41	58.95	1442.23	2573.9919360000004	12.4
128	142.196	90.905	14.989	13.153	56.96	1416.58	2446.065664	11.9
120	131.896	85.815	15.253	13.072	52.124	1419.31	3066.560512000001	12.3
112	127.43	80.222	52.491	9.829	116.563	1362.43	2643.7222400000005	9.7
104	113.591	74.318	14.549	9.401	46.07	1412.72	2236.3504639999996	12.3

96	110.533	70.568	14.725	8.685	43.051	1361.89	2034.4995839999995	11.4

The model config "resnet50_config_2" uses 4 GPU instances with a max batch size of 128 and has dynamic batching enabled. 21 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.