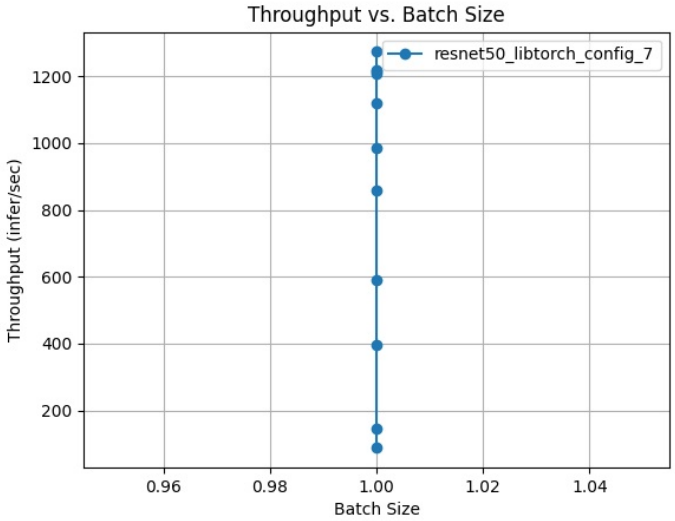
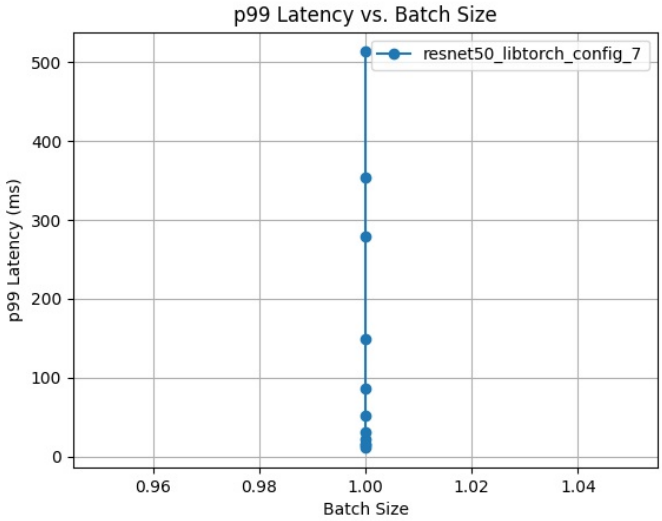
Detailed Report

Model Config: resnet50_libtorch_config_7





Throughput vs. Batch Size curves for config resnet50_libtorch_config_7 **p99 Latency vs. Batch Size curves for config resnet50_libtorch_config_7**

Client Batch Size	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 (%)
1	353.539	305.309	130.354	20.89	121.295	1273.38	5598.347264	94.3
1	513.467	413.317	231.214	20.686	126.777	1217.53	10264.510464	94.7
1	148.785	105.458	24.143	8.732	56.894	1215.41	10222.567424	64.0
1	279.885	207.572	63.45	18.291	99.944	1206.96	10285.481984	62.7
1	85.611	56.625	7.689	5.272	35.794	1120.12	10201.595904	72.0
1	52.374	32.239	3.996	2.505	22.214	985.962	5942.280192	56.0
1	30.184	18.475	2.189	0.982	13.786	859.755	3671.064576	70.0
1	22.117	13.385	0.38	0.243	13.414	590.318	2674.917376	55.7
1	15.654	9.988	0.05	0.218	8.939	396.438	2674.917376	78.2
1	10.879	6.748	0.09	0.195	5.61	145.876	2662.334464	34.0
1	15.023	11.212	0.106	0.248	13.055	87.8018	2662.334464	27.3

The model config "resnet50_libtorch_config_7" uses 6 GPU instances with a max batch size of 32 and has dynamic batching enabled. 11 measurement(s) were obtained for the model config on GPU(s) 1 x NVIDIA TITAN RTX with total memory 23.7 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.