Online Result Summary

Model: resnet50_libtorch

GPU(s): 1 x NVIDIA TITAN RTX

Total Available GPU Memory: 23.7 GB

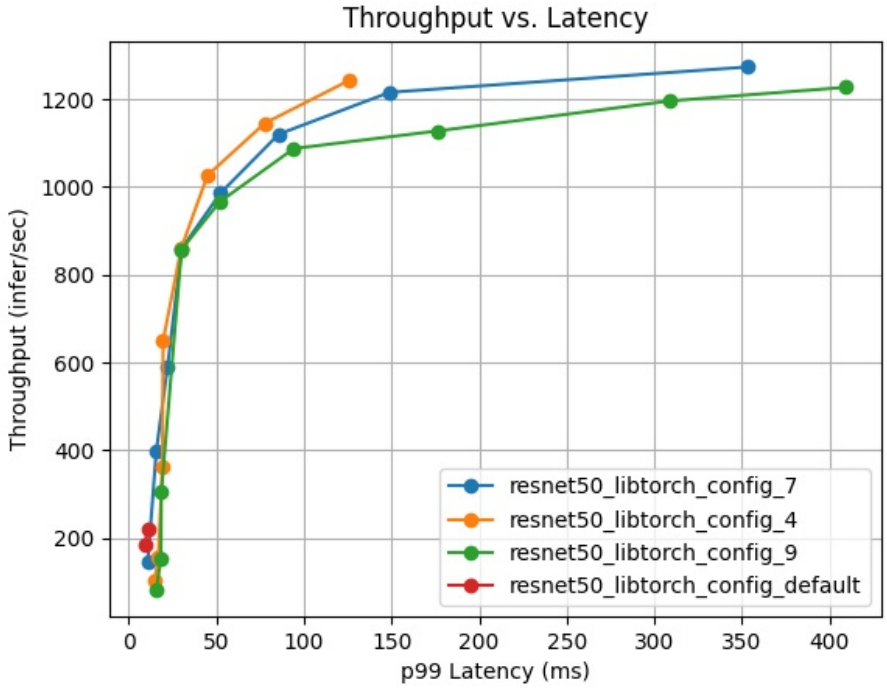
Constraint targets: None

In 46 measurements across 12 configurations, **resnet50_libtorch_config_7** provides the best throughput: **1273 infer/sec**.

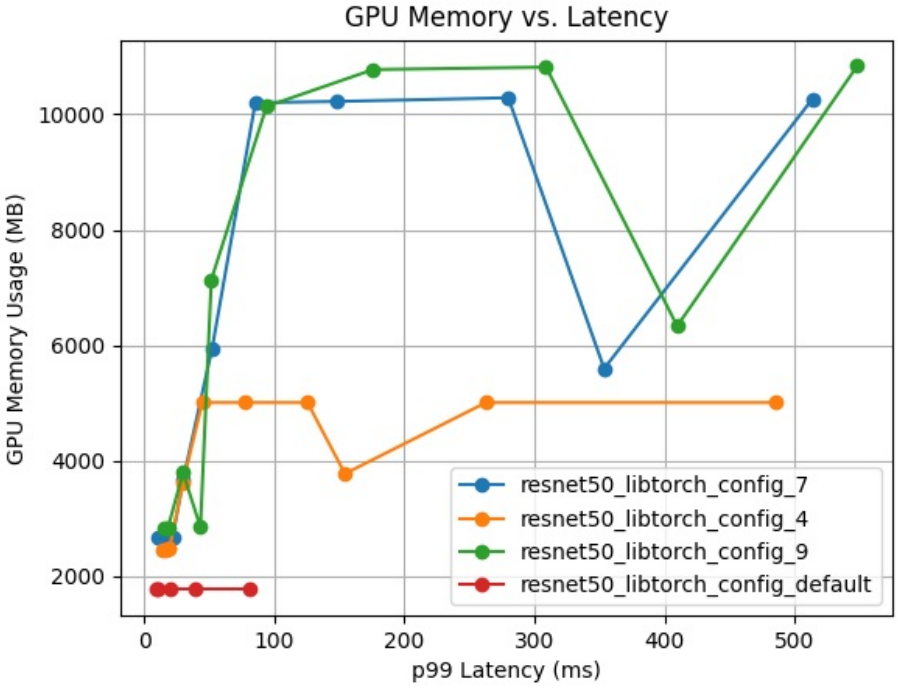
This is a **473% gain** over the default configuration (222 infer/sec), under the given constraints on GPU(s) 1 x NVIDIA TITAN RTX.

- **resnet50_libtorch_config_7**: 6 GPU instances with a max batch size of 32 on platform pytorch_libtorch

Curves corresponding to the 3 best model configuration(s) out of a total of 12 are shown in the plots.



Throughput vs. Latency curves for 3 best configurations.



GPU Memory vs. Latency curves for 3 best configurations.

The following table summarizes each configuration at the measurement that optimizes the desired metrics under the given constraints.

Model Config Name	Max Batch Size	Dynamic Batching	Total Instance Count	p99 Latency (ms)	Throughput (infer/sec)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
resnet50_libtorch_config_7	32	Enabled	6:GPU	353.539	1273.38	5598	94.3
resnet50_libtorch_config_4	16	Enabled	5:GPU	125.635	1242.25	5011	90.2
resnet50_libtorch_config_9	32	Enabled	7:GPU	409.616	1227.02	6334	95.3
resnet50_libtorch_config_default	128	Disabled	1:GPU	10.938	221.846	1773	41.0