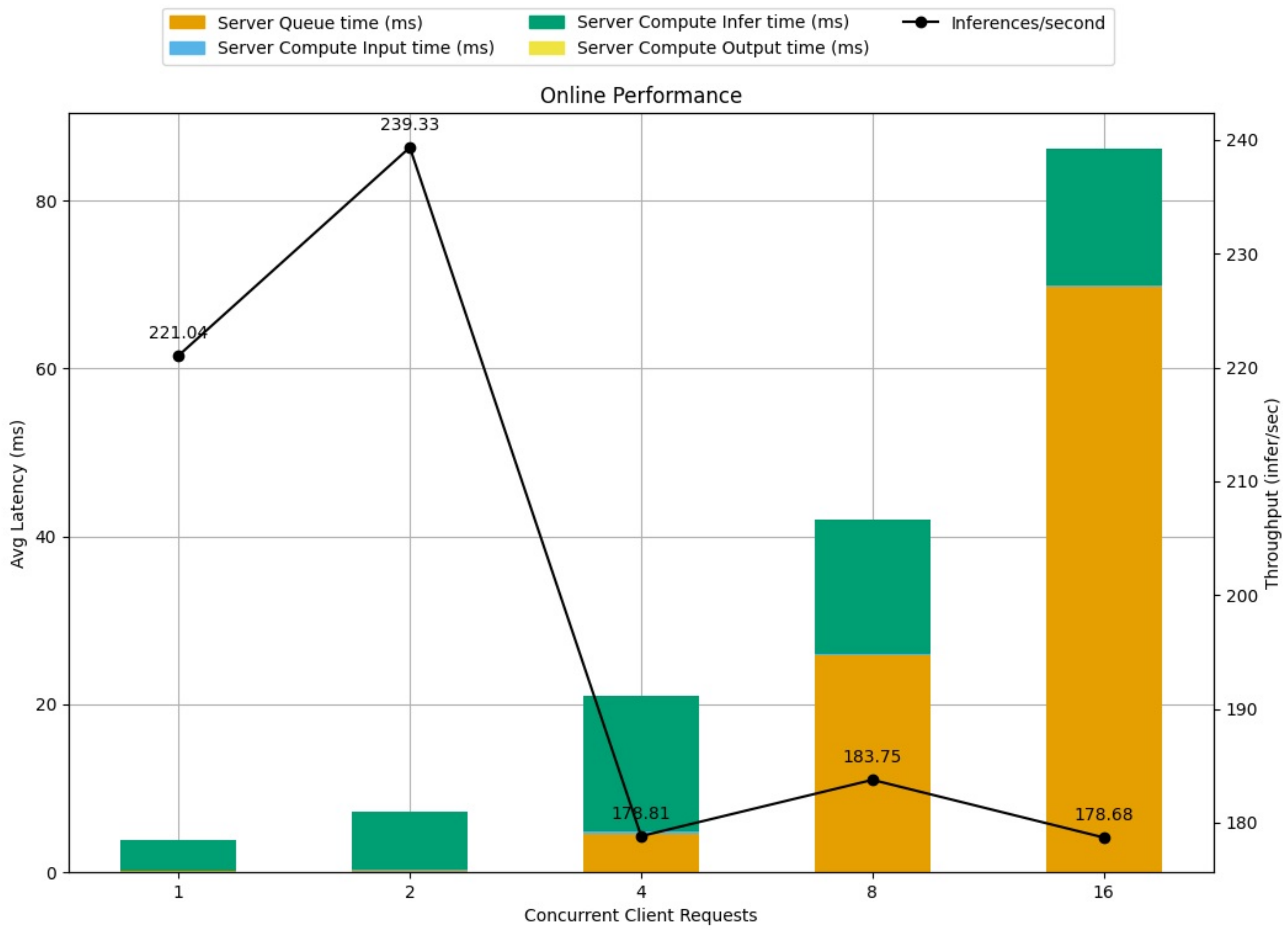
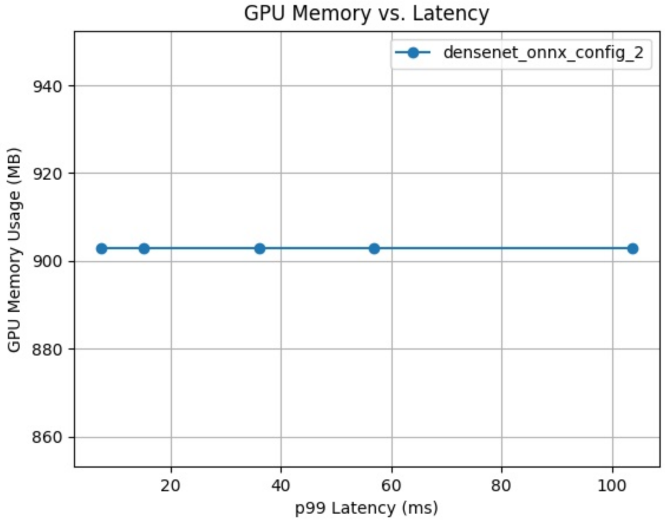
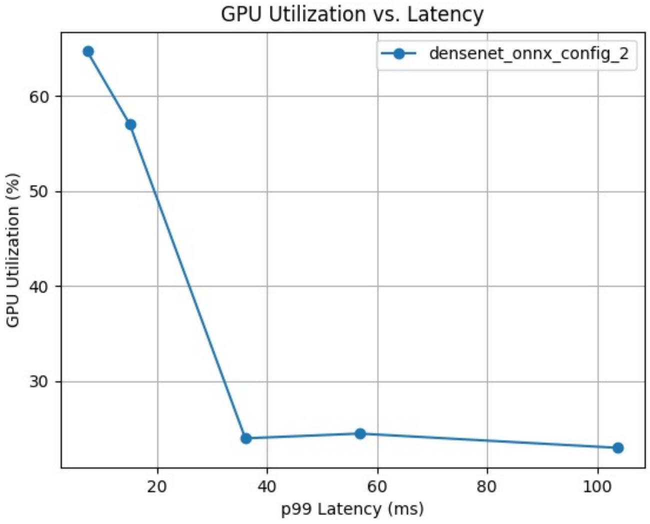
Detailed Report

Model Config: densenet_onnx_config_2



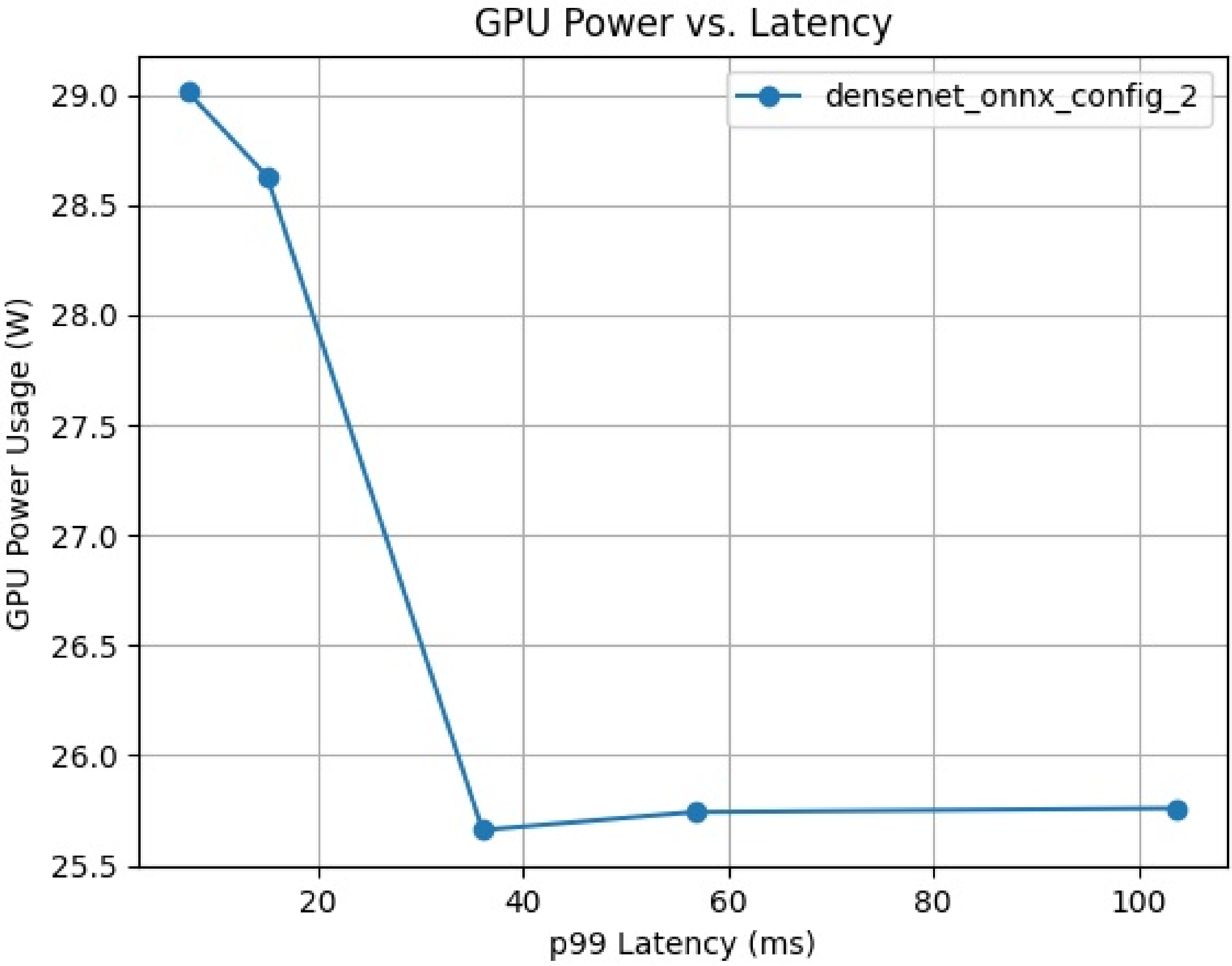
Latency Breakdown for Online Performance of densenet_onnx_config_2





GPU Memory vs. Latency curves for config densenet_onnx_config_2

GPU Utilization vs. Latency curves for config densenet_onnx_config_2



GPU Power vs. Latency curves for config densenet_onnx_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 (%)
16	103.803	87.388	69.688	0.177	16.265	178.683	902.823936	23.0
8	56.776	43.015	25.895	0.169	15.908	183.753	902.823936	24.5
4	36.071	22.077	4.611	0.186	16.197	178.806	902.823936	24.0
2	15.079	8.206	0.166	0.126	6.882	239.332	902.823936	57.0
1	7.355	4.417	0.111	0.093	3.61	221.039	902.823936	64.7

The model config **densenet_onnx_config_2** uses 3 GPU instances with a max batch size of 0 and has dynamic batching enabled. 5 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 onnxruntime_onnx.

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.