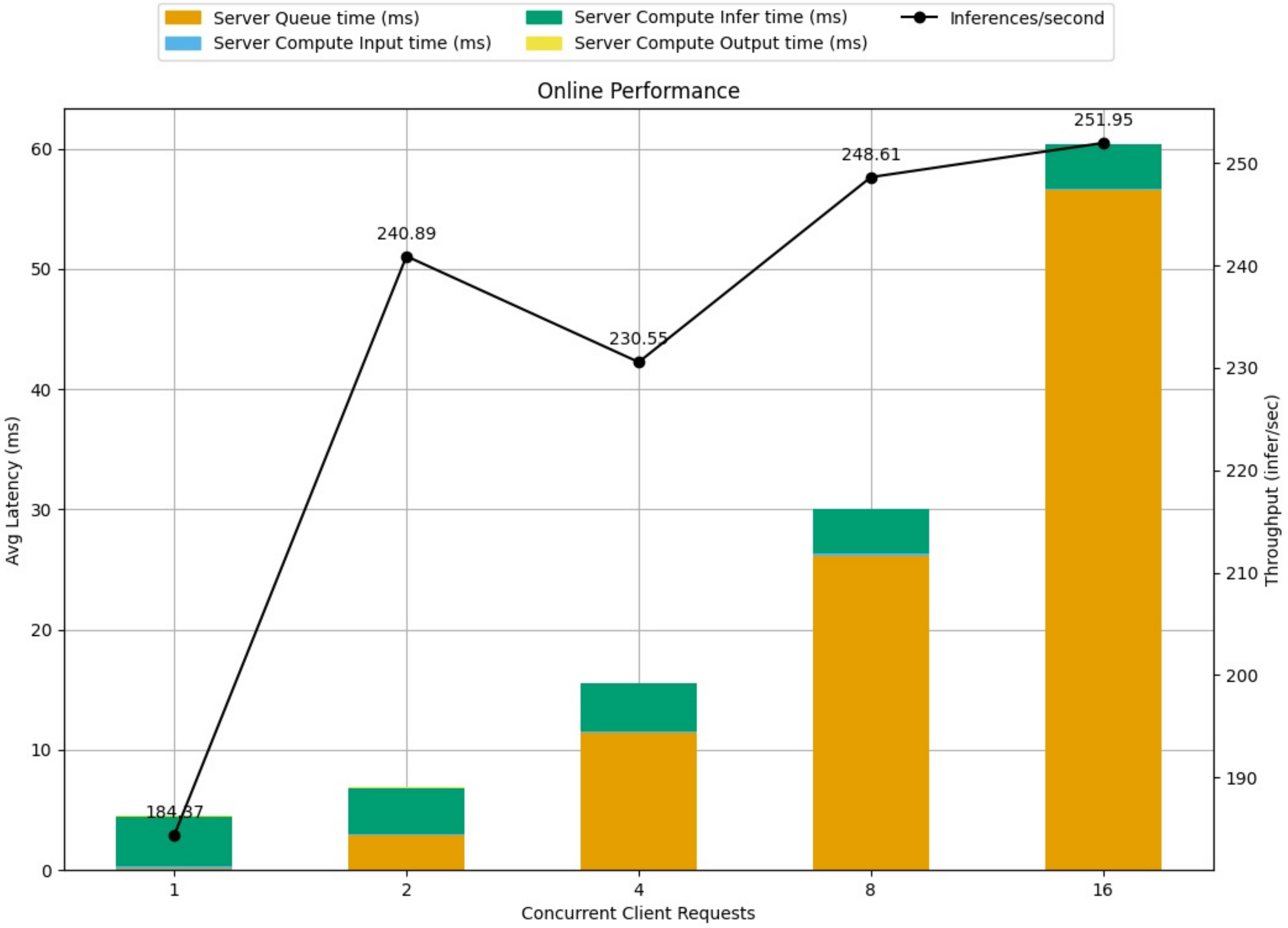
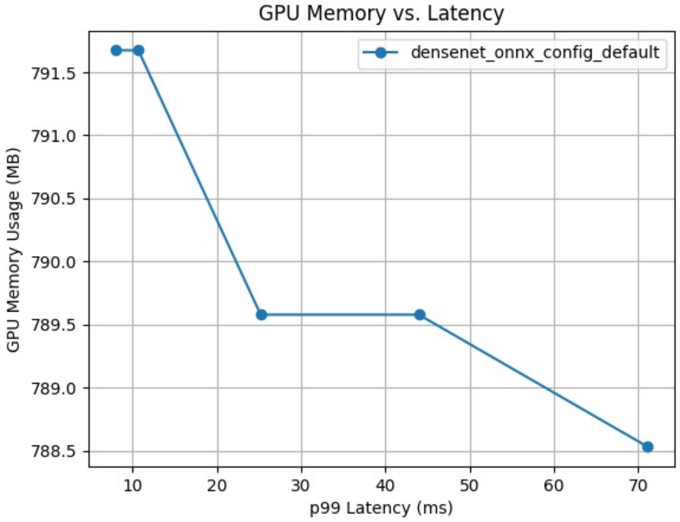
Detailed Report

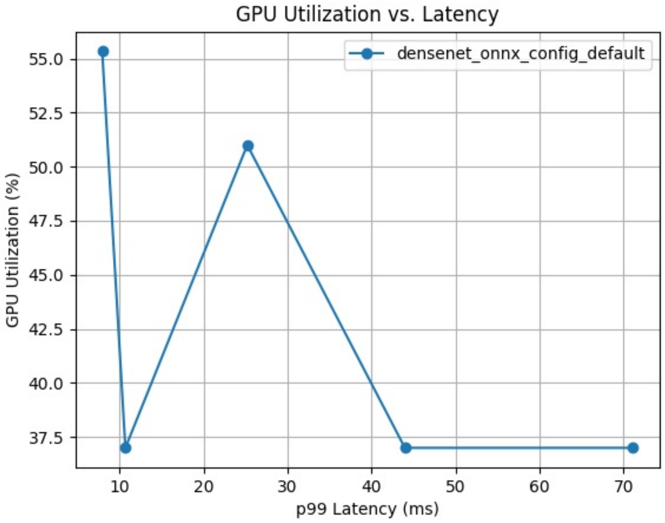
Model Config: densenet_onnx_config_default



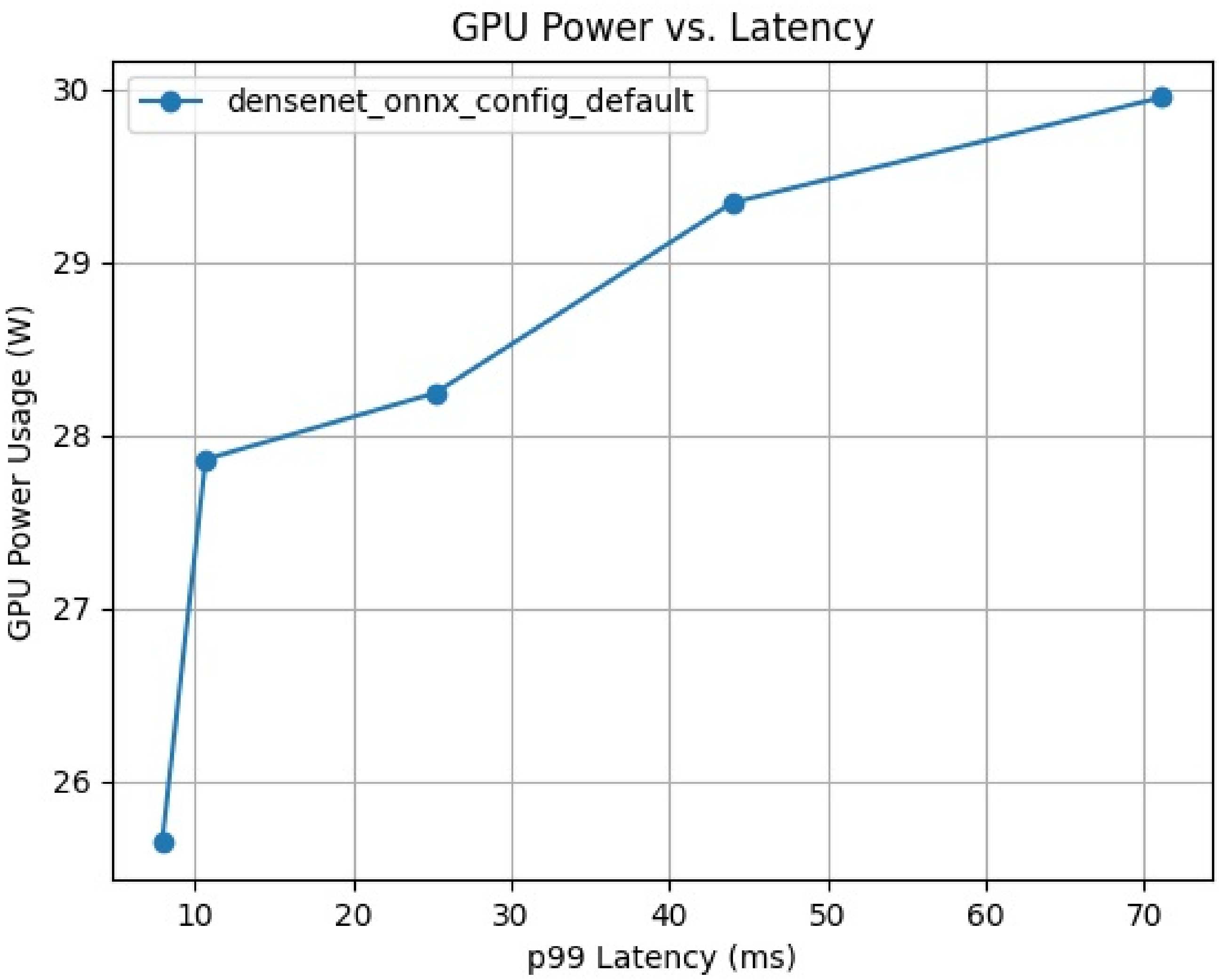
Latency Breakdown for Online Performance of densenet_onnx_config_default



GPU Memory vs. Latency curves for config densenet_onnx_config_default



GPU Utilization vs. Latency curves for config densenet_onnx_config_default



GPU Power vs. Latency curves for config densenet_onnx_config_default

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	71.122	61.853	56.549	0.109	3.676	251.949	788.529152	37.0
8	43.951	31.498	26.175	0.111	3.737	248.613	789.577728	37.0
4	25.184	16.928	11.383	0.123	4.021	230.554	789.577728	51.0
2	10.671	8.086	2.911	0.111	3.818	240.893	791.67488	37.0
1	7.914	5.271	0.152	0.117	4.188	184.365	791.67488	55.3

The model config **densenet_onnx_config_default** uses 1 GPU instance 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.