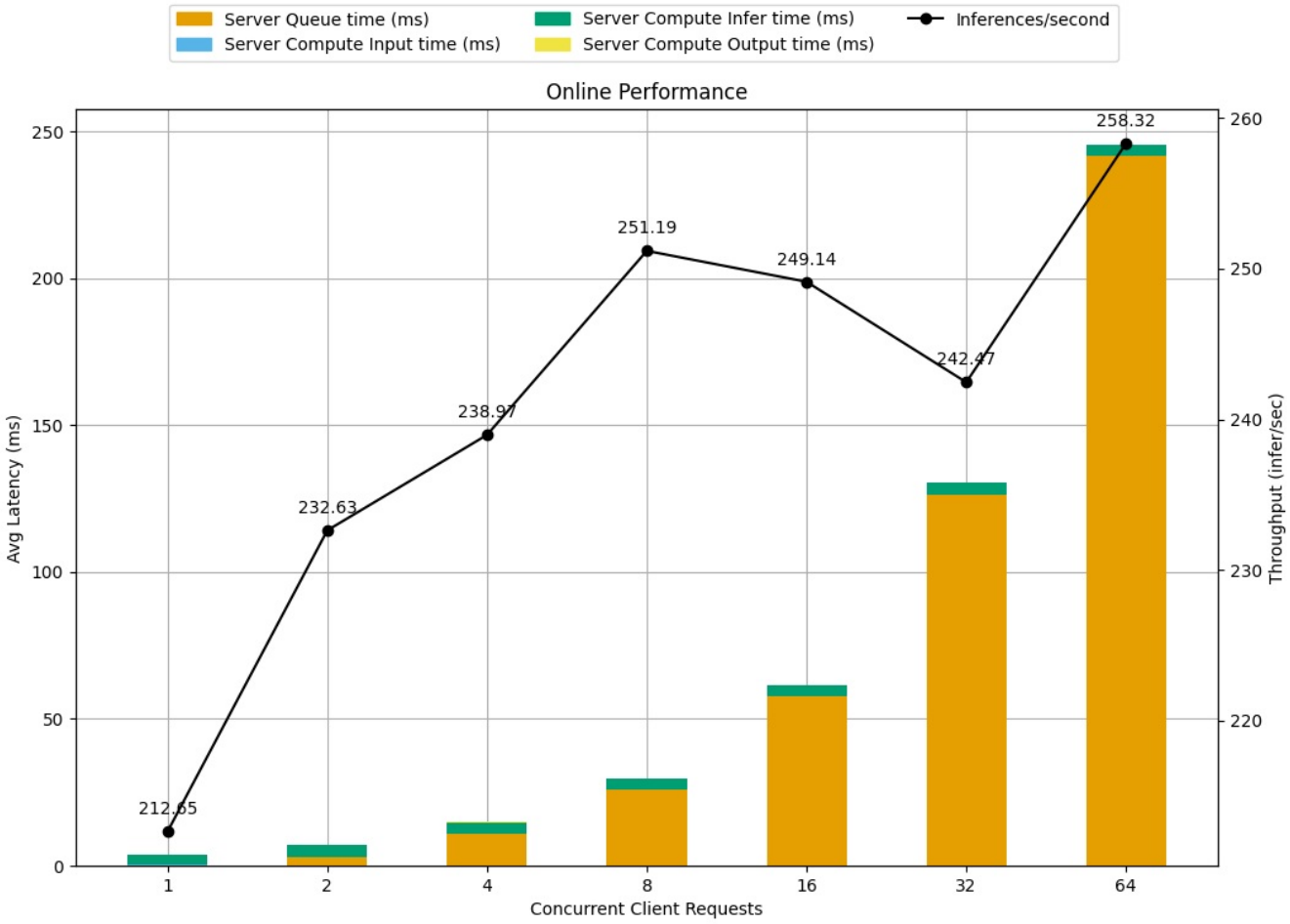
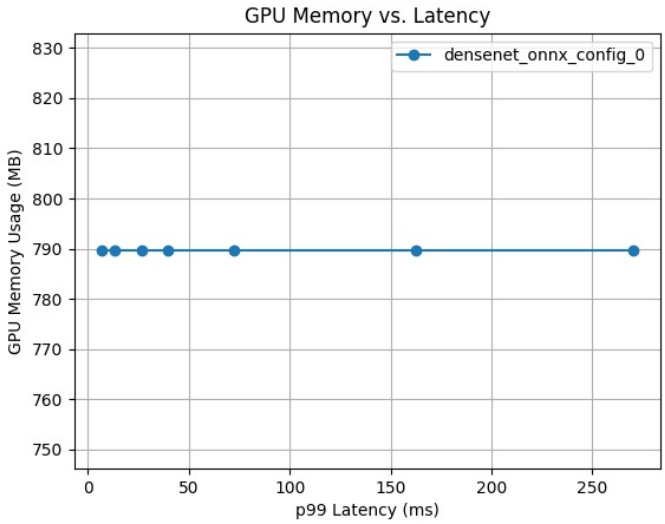
Detailed Report

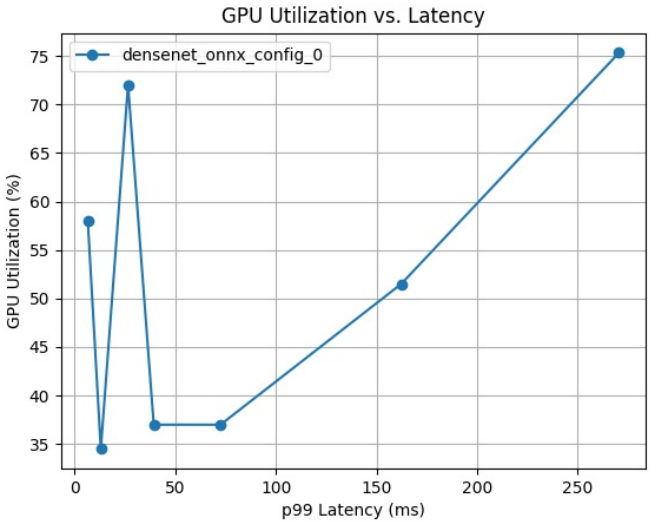
Model Config: densenet_onnx_config_0



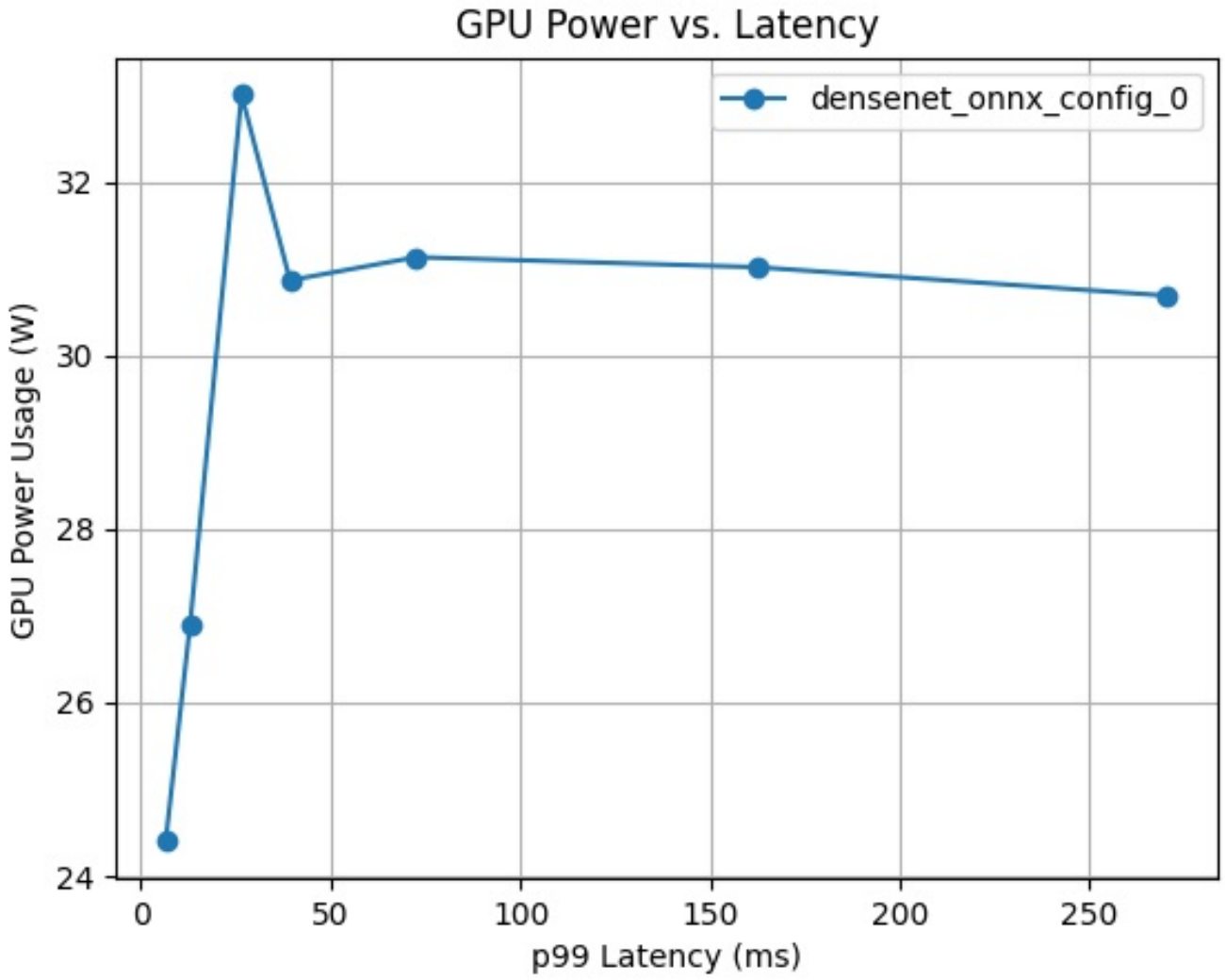
Latency Breakdown for Online Performance of densenet_onnx_config_0



GPU Memory vs. Latency curves for config densenet_onnx_config_0



GPU Utilization vs. Latency curves for config densenet_onnx_config_0



GPU Power vs. Latency curves for config densenet_onnx_config_0

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 (%)
64	270.471	246.703	241.593	0.106	3.618	258.316	789.577728	75.3
32	162.24	131.907	126.357	0.11	3.855	242.465	789.577728	51.5
16	72.48	62.803	57.5	0.121	3.697	249.136	789.577728	37.0
8	39.29	31.144	25.951	0.118	3.685	251.19	789.577728	37.0
4	26.526	16.217	10.813	0.105	3.921	238.972	789.577728	72.0
2	12.989	8.363	2.898	0.11	3.908	232.635	789.577728	34.5
1	6.525	4.563	0.135	0.104	3.609	212.649	789.577728	58.0

The model config **densenet_onnx_config_0** uses 1 GPU instance with a max batch size of 0 and has dynamic batching enabled. 7 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.