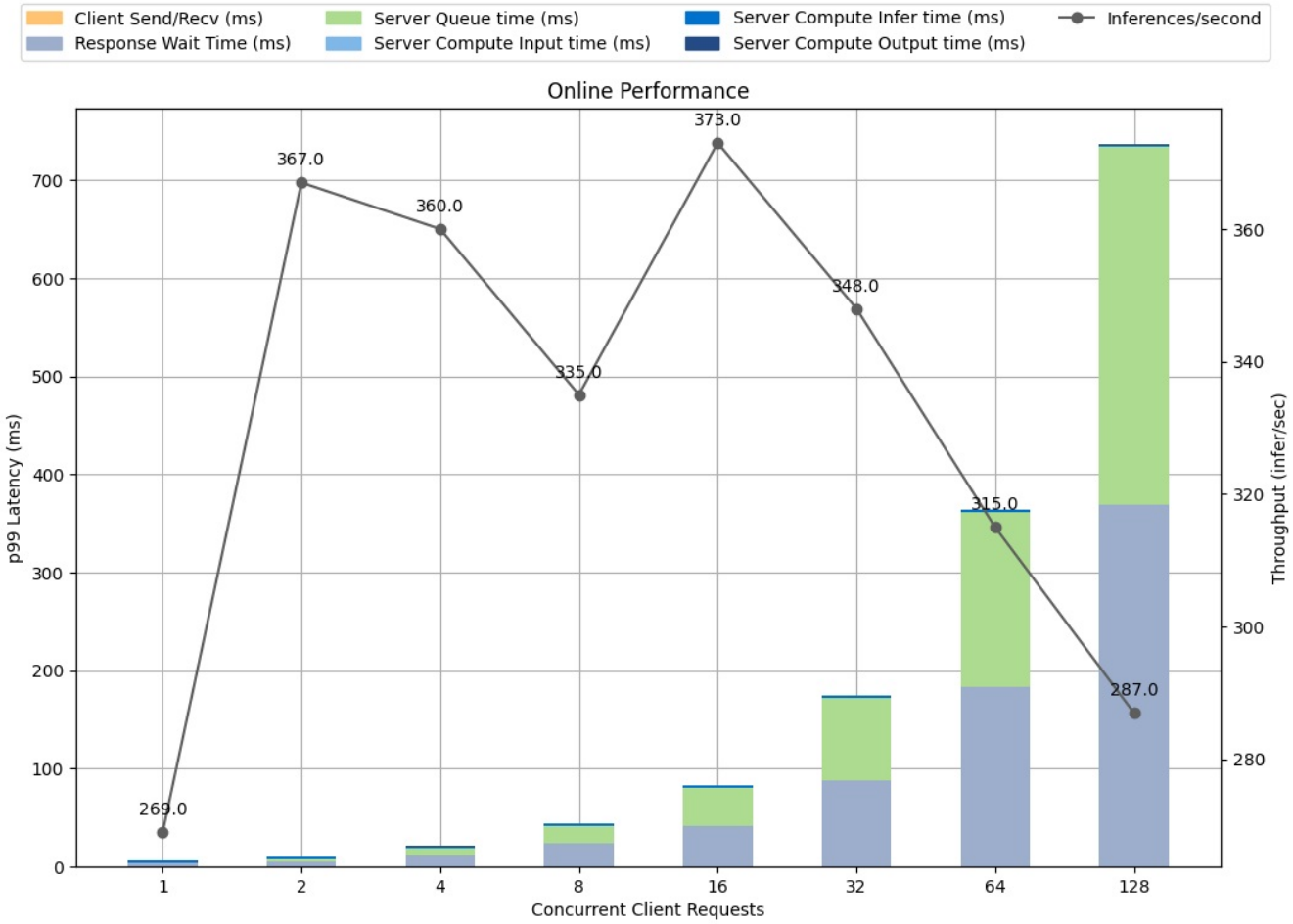
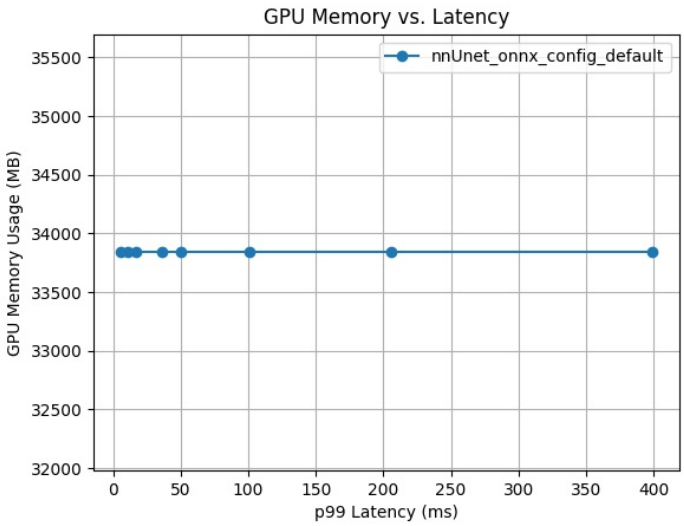
Detailed Report

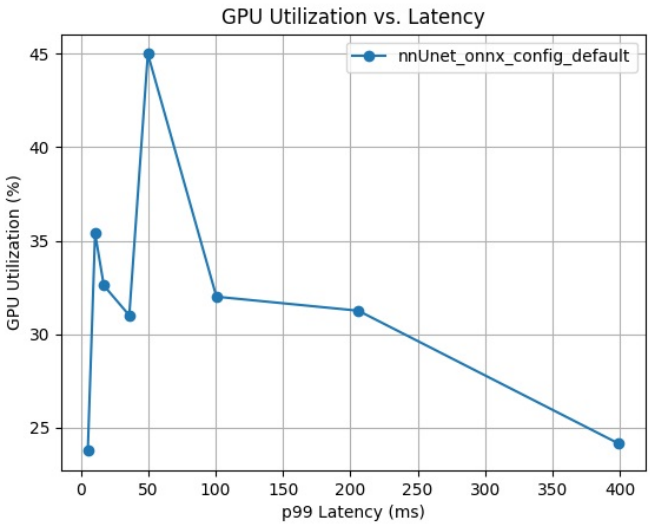
Model Config: nnUnet_onnx_config_default



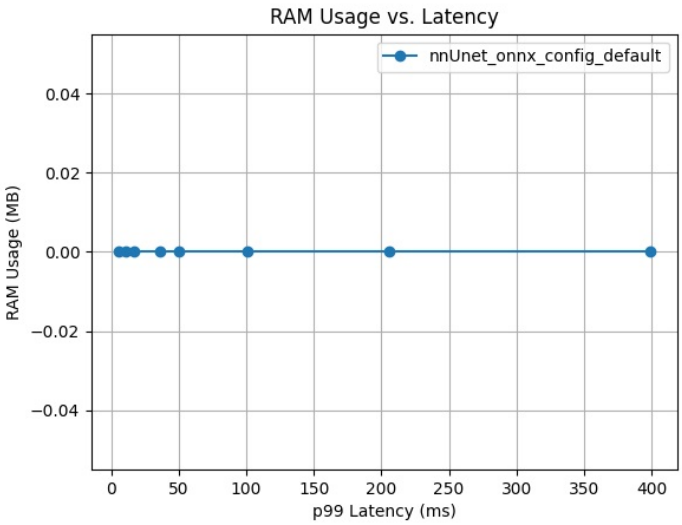
Latency Breakdown for Online Performance of nnUnet_onnx_config_default



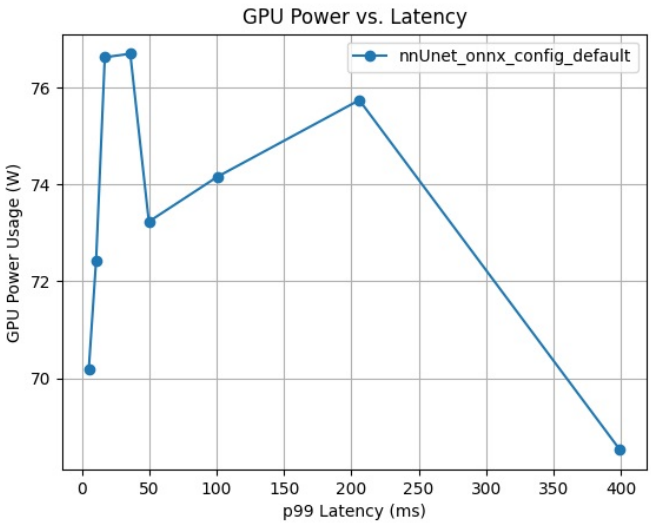
GPU Memory vs. Latency curves for config nnUnet_onnx_config_default



GPU Utilization vs. Latency curves for config nnUnet_onnx_config_default



RAM Usage vs. Latency curves for config nnUnet_onnx_config_default



GPU Power vs. Latency curves for config nnUnet_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 CPU Memory Usage (MB)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
128	399.304	369.079	364.803	0.178	2.126	287.0	0	33839.0	24.1
64	205.882	182.598	178.53	0.13	2.17	315.0	0	33839.0	31.2
32	100.625	87.823	83.81	0.129	2.138	348.0	0	33839.0	32.0
16	49.762	41.919	38.2	0.111	2.09	373.0	0	33839.0	45.0
8	35.972	23.551	17.486	0.125	2.337	335.0	0	33839.0	31.0
4	17.095	10.961	7.239	0.117	2.169	360.0	0	33839.0	32.6
2	10.618	5.354	1.869	0.12	2.183	367.0	0	33839.0	35.4
1	5.218	3.609	0.041	0.122	2.136	269.0	0	33839.0	23.8

The model config "nnUnet_onnx_config_default" uses 1 GPU instance(s) with a max batch size of 1 and has dynamic batching enabled. 8 measurement(s) were obtained for the model config on GPU(s) NVIDIA A100-PCIE-40GB with memory limit(s) 39.4 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 throughput.