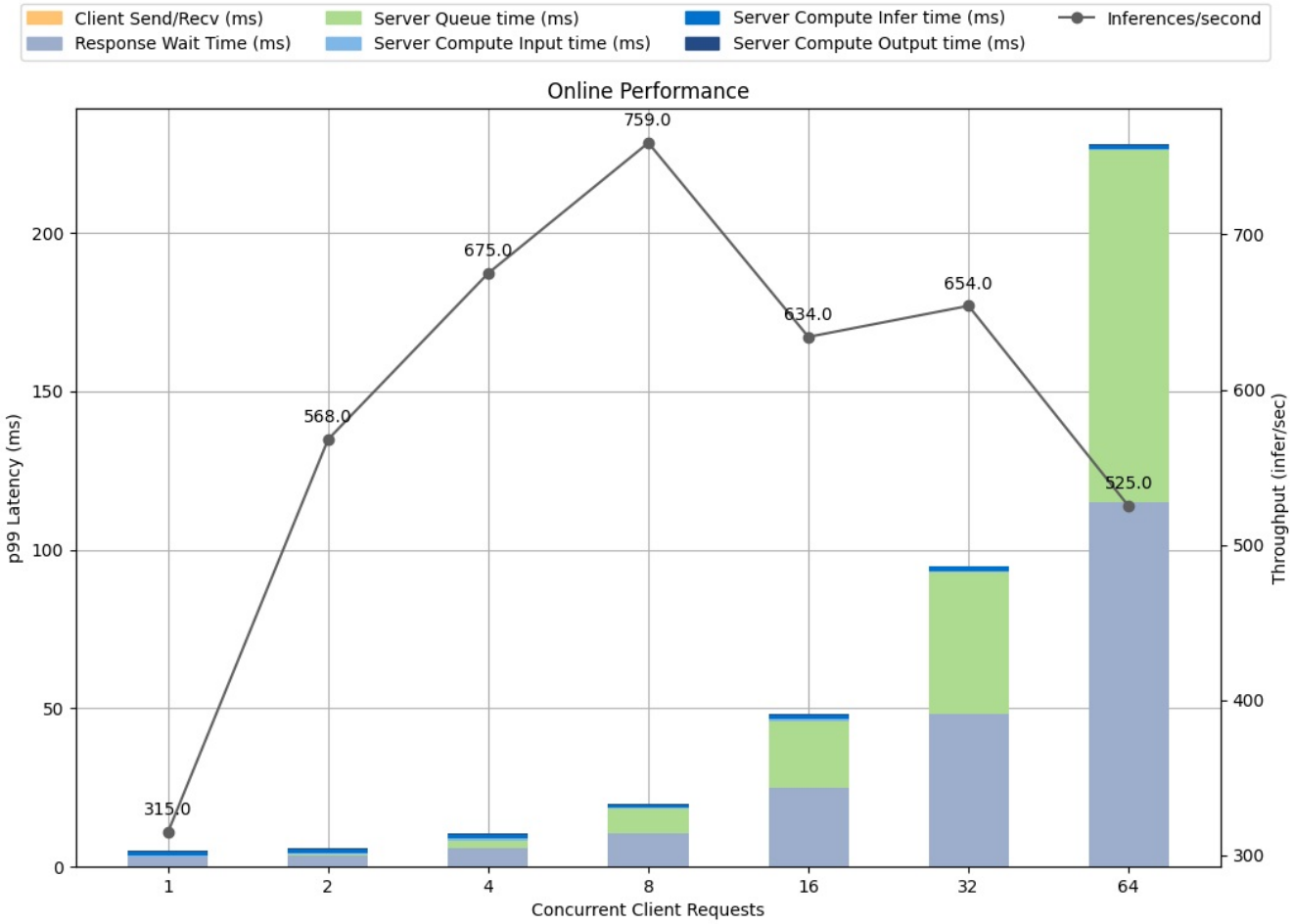
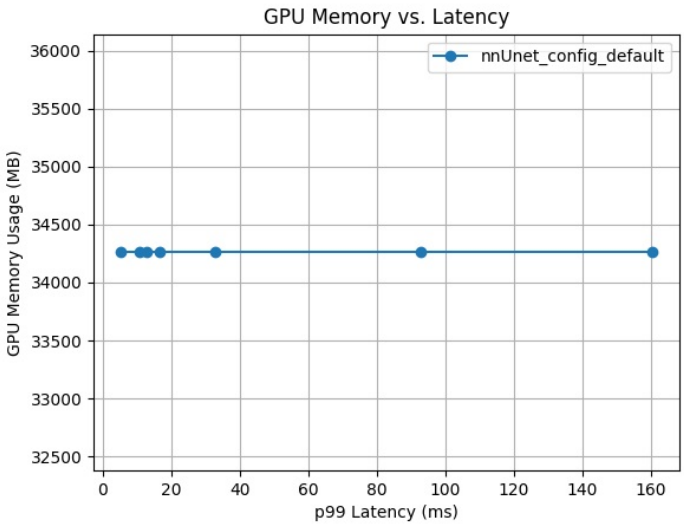
Detailed Report

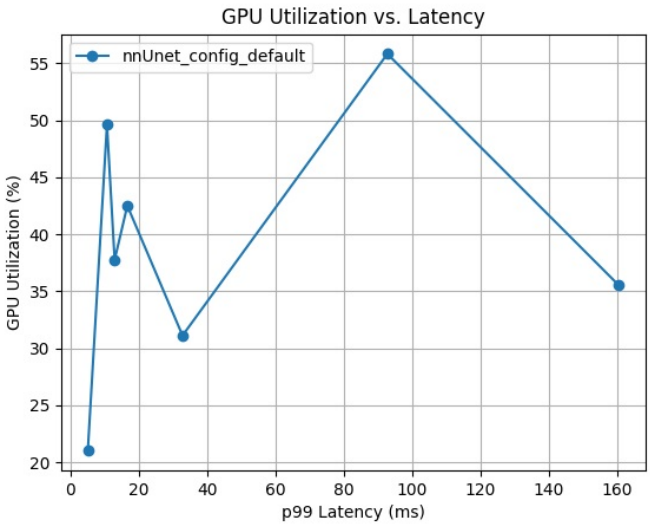
Model Config: nnUnet_config_default



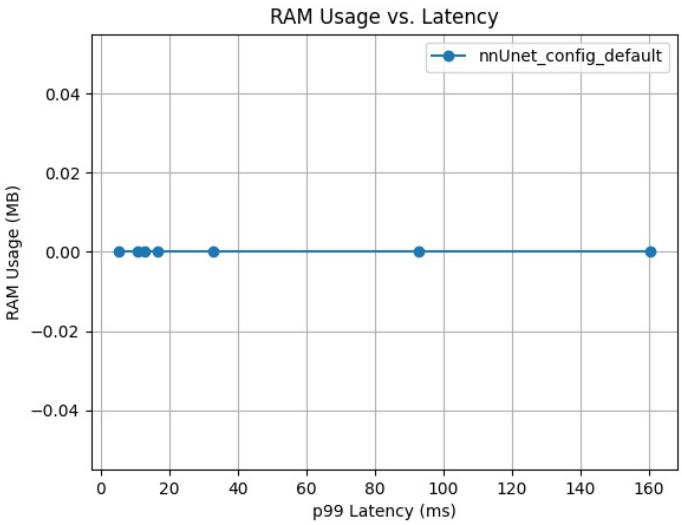
Latency Breakdown for Online Performance of nnUnet_config_default



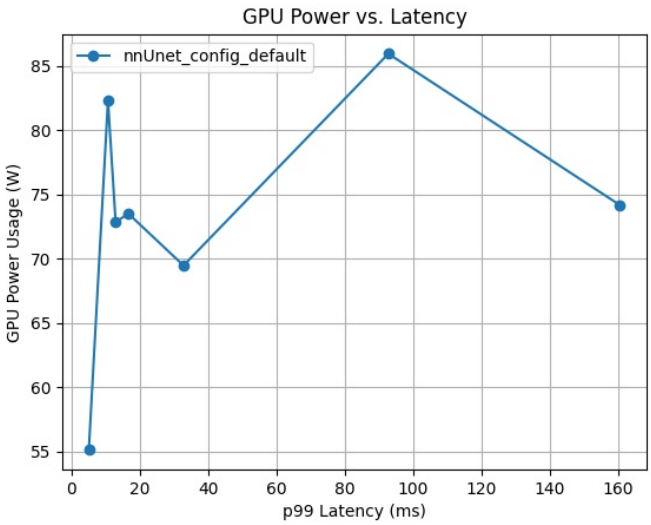
GPU Memory vs. Latency curves for config nnUnet_config_default



GPU Utilization vs. Latency curves for config nnUnet_config_default



RAM Usage vs. Latency curves for config nnUnet_config_default



GPU Power vs. Latency curves for config nnUnet_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 (%)
64	160.573	114.785	111.031	0.468	1.313	525.0	0	34261.0	35.5
32	92.792	48.17	44.444	0.685	1.272	654.0	0	34261.0	55.8
16	32.861	24.638	21.091	0.664	1.361	634.0	0	34261.0	31.1
8	16.709	10.377	7.674	0.372	1.096	759.0	0	34261.0	42.5
2	12.998	3.434	0.421	0.446	1.247	568.0	0	34261.0	37.8
4	10.758	5.803	2.451	0.615	1.296	675.0	0	34261.0	49.6
1	5.148	3.08	0.043	0.274	1.366	315.0	0	34261.0	21.0

The model config "nnUnet_config_default" uses 1 GPU instance(s) with a max batch size of 1 and has dynamic batching enabled. 7 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 tensorrt_plan.

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.