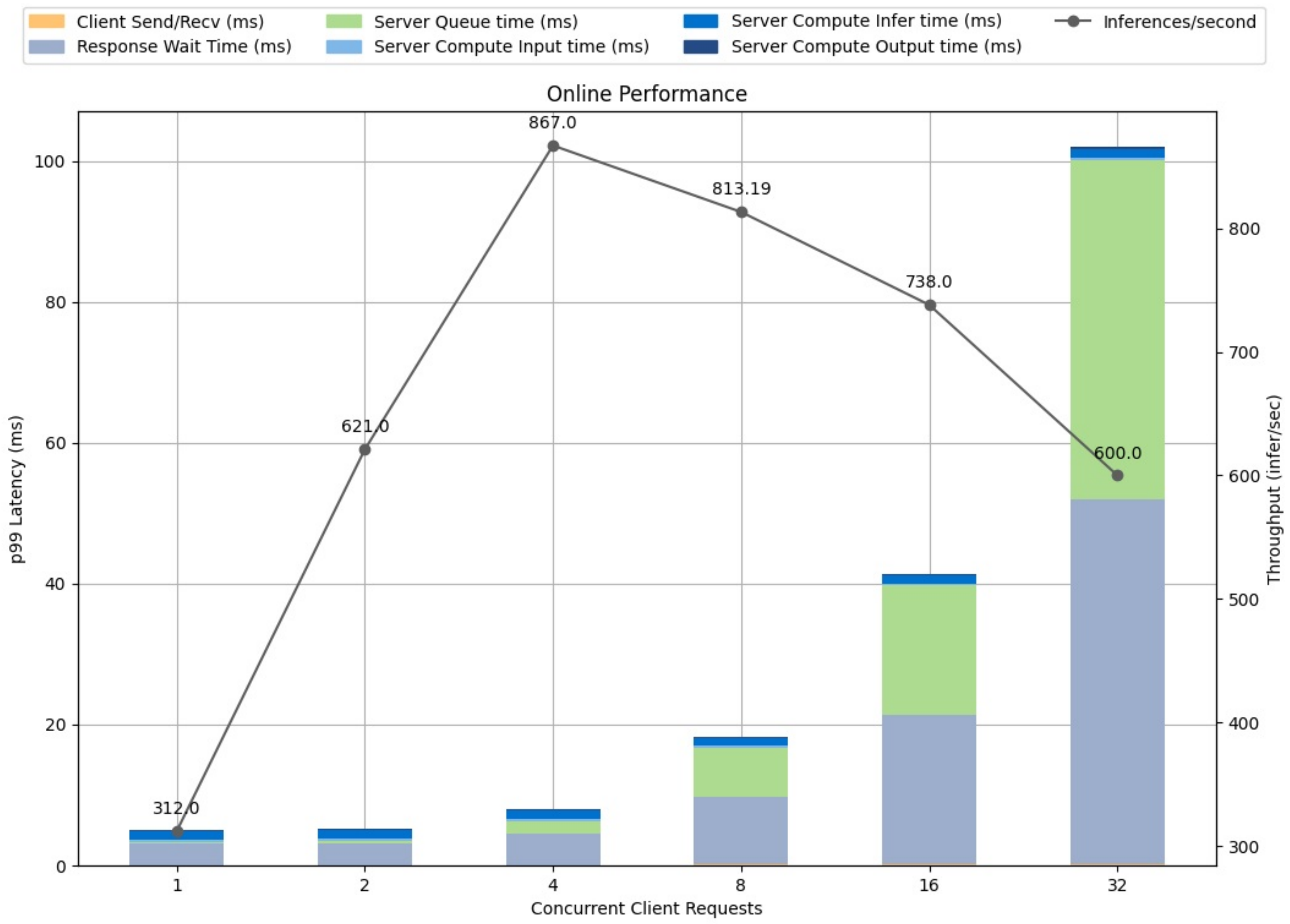
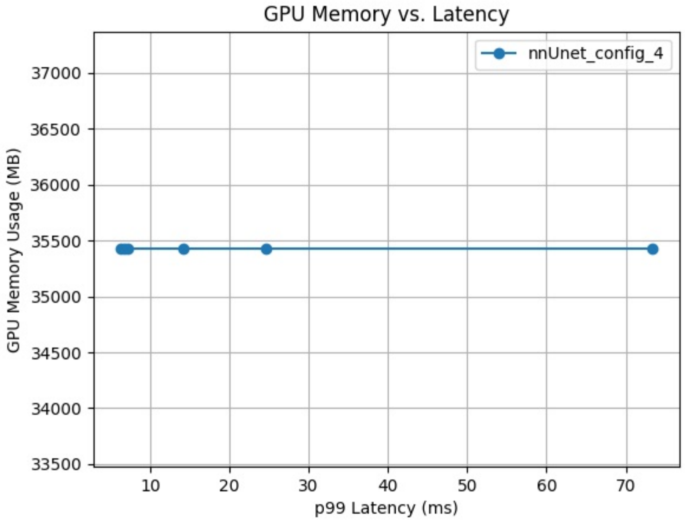
Detailed Report

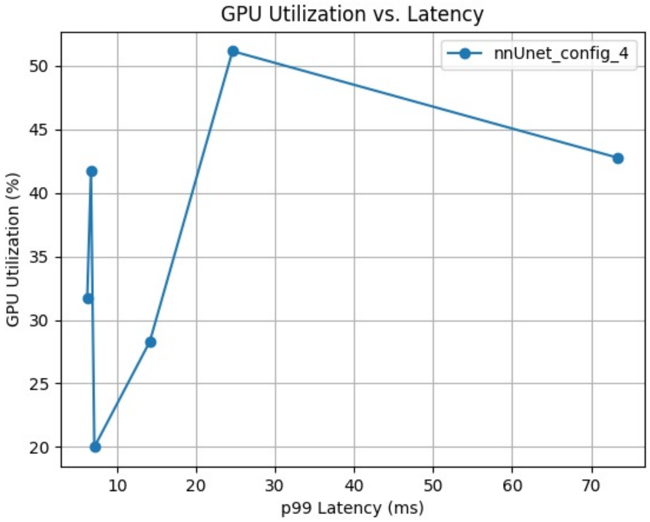
Model Config: nnUnet_config_4



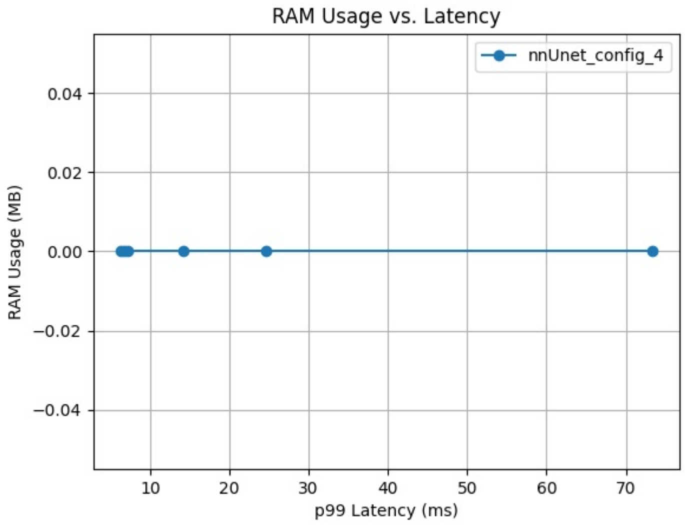
Latency Breakdown for Online Performance of nnUnet_config_4



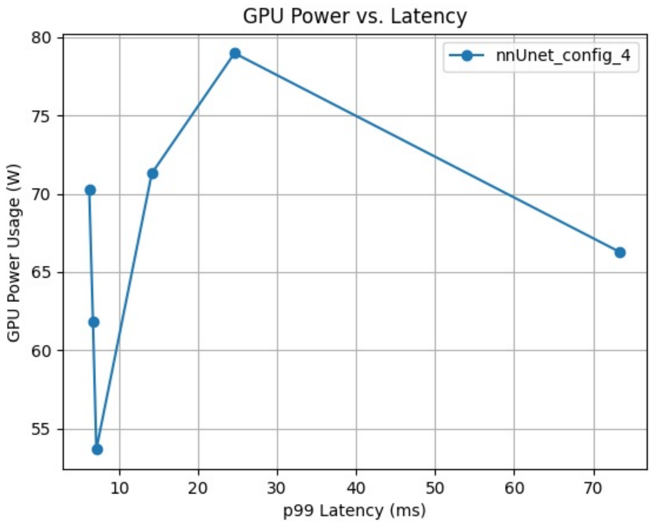
GPU Memory vs. Latency curves for config nnUnet_config_4



GPU Utilization vs. Latency curves for config nnUnet_config_4



RAM Usage vs. Latency curves for config nnUnet_config_4



GPU Power vs. Latency curves for config nnUnet_config_4

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 (%)
32	73.409	51.79	48.178	0.349	1.237	600.0	0	35424.0	42.8
16	24.614	21.31	18.377	0.265	1.116	738.0	0	35424.0	51.1
8	14.145	9.684	6.954	0.273	1.125	813.187	0	35424.0	28.3
1	7.13	3.119	0.094	0.302	1.289	312.0	0	35424.0	20.0
4	6.702	4.512	1.714	0.301	1.18	867.0	0	35424.0	41.7
2	6.229	3.125	0.315	0.293	1.201	621.0	0	35424.0	31.8

The model config "nnUnet_config_4" uses 5 GPU instance(s) with a max batch size of 1 and has dynamic batching enabled. 6 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.