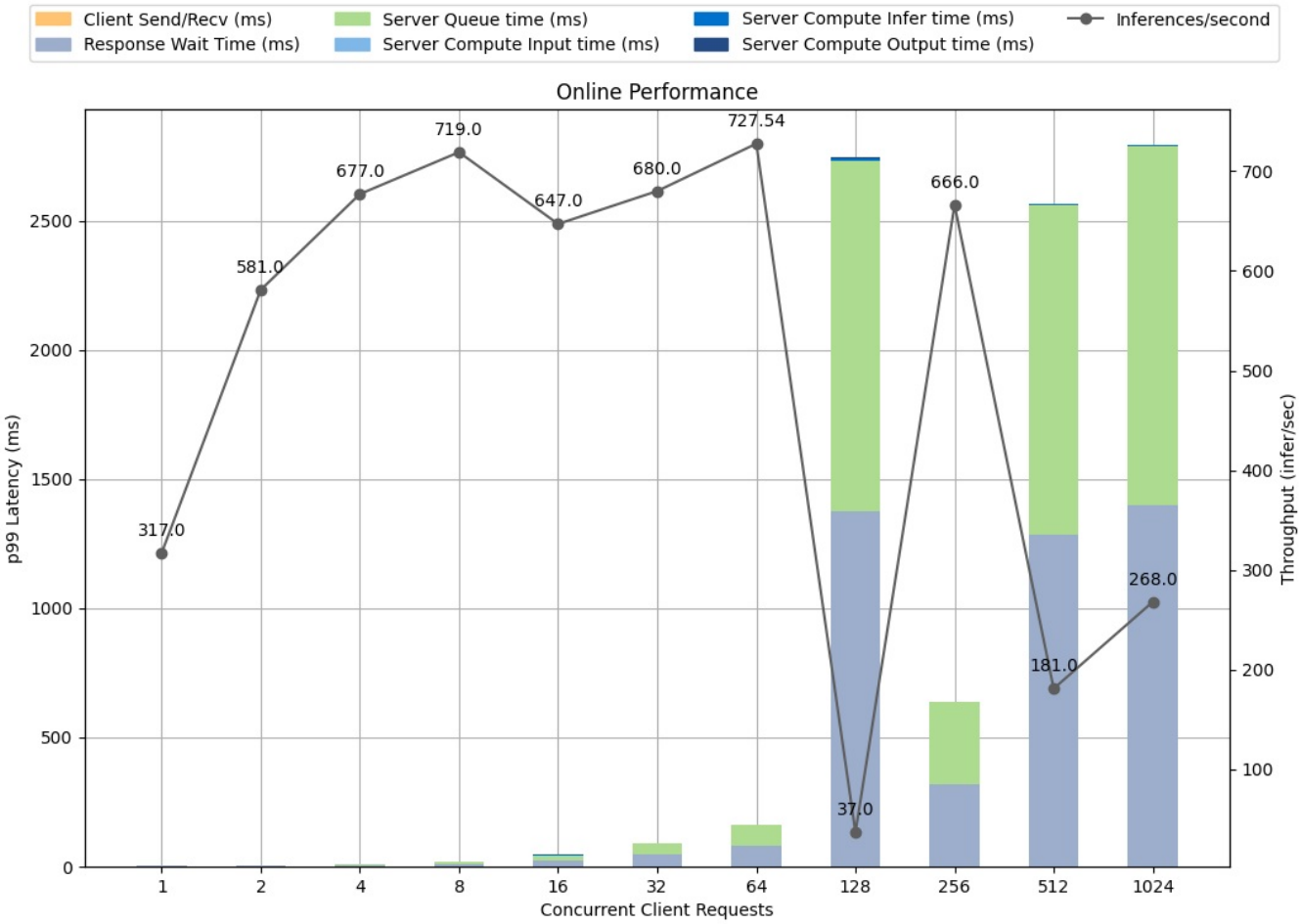
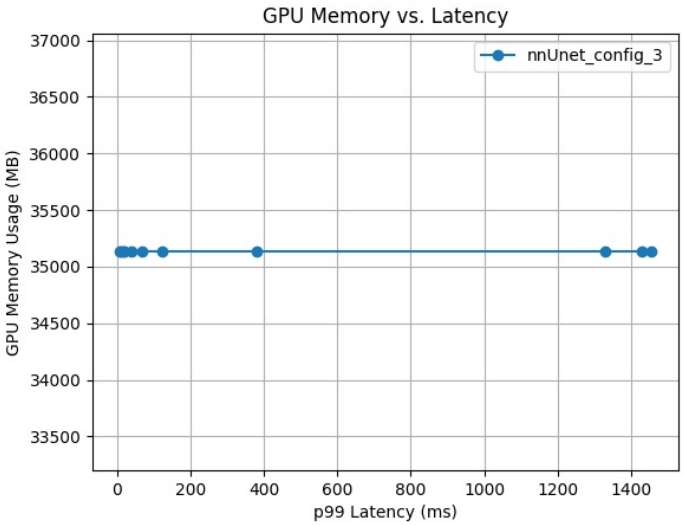
Detailed Report

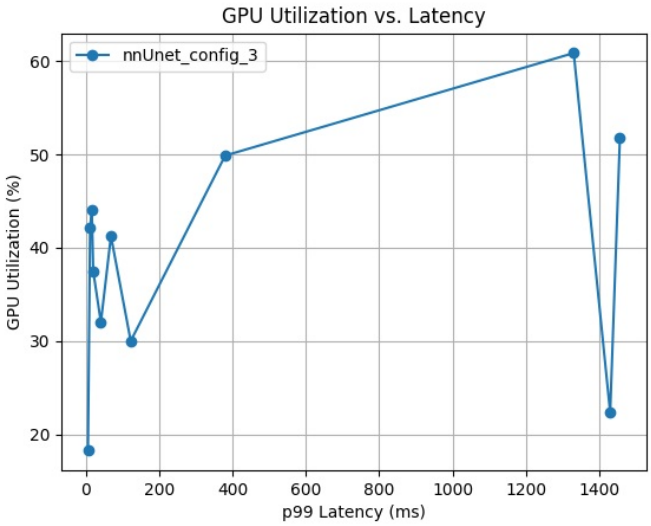
Model Config: nnUnet_config_3



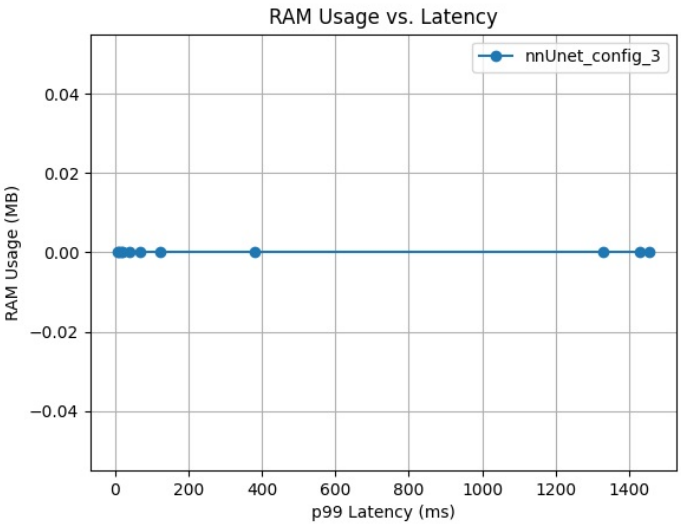
Latency Breakdown for Online Performance of nnUnet_config_3



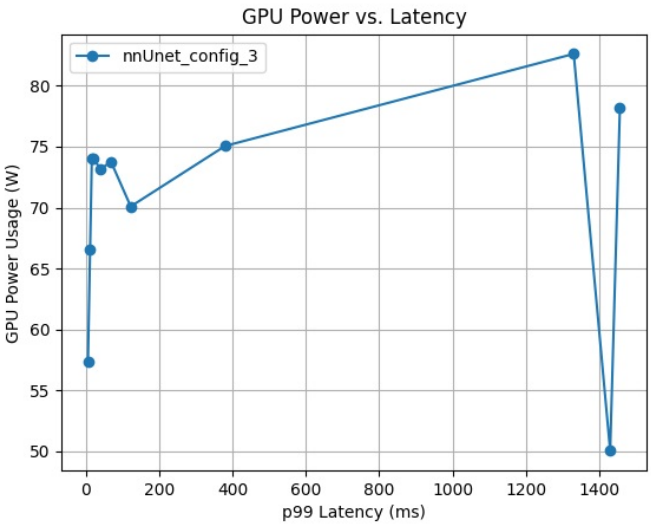
GPU Memory vs. Latency curves for config nnUnet_config_3



GPU Utilization vs. Latency curves for config nnUnet_config_3



RAM Usage vs. Latency curves for config nnUnet_config_3



GPU Power vs. Latency curves for config nnUnet_config_3

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 (%)
512	1456.454	1282.92	1278.55	0.505	1.567	181.0	0	35131.0	51.7
128	1429.951	1377.008	1354.027	0.868	9.81	37.0	0	35131.0	22.3
1024	1330.545	1396.608	1393.139	0.322	1.313	268.0	0	35131.0	60.9
256	379.798	320.006	316.831	0.345	1.186	666.0	0	35131.0	49.9
64	122.132	81.861	78.37	0.397	1.203	727.545	0	35131.0	30.0
32	68.212	45.804	42.71	0.346	1.157	680.0	0	35131.0	41.2
16	40.744	24.17	19.674	0.729	1.166	647.0	0	35131.0	32.0
4	19.863	5.79	2.335	0.459	1.17	677.0	0	35131.0	37.5
8	16.08	10.954	8.046	0.321	1.12	719.0	0	35131.0	44.0
2	11.374	3.385	0.37	0.425	1.139	581.0	0	35131.0	42.2
1	5.529	3.058	0.041	0.275	1.403	317.0	0	35131.0	18.2

The model config "nnUnet_config_3" uses 4 GPU instance(s) with a max batch size of 1 and has dynamic batching enabled. 11 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.