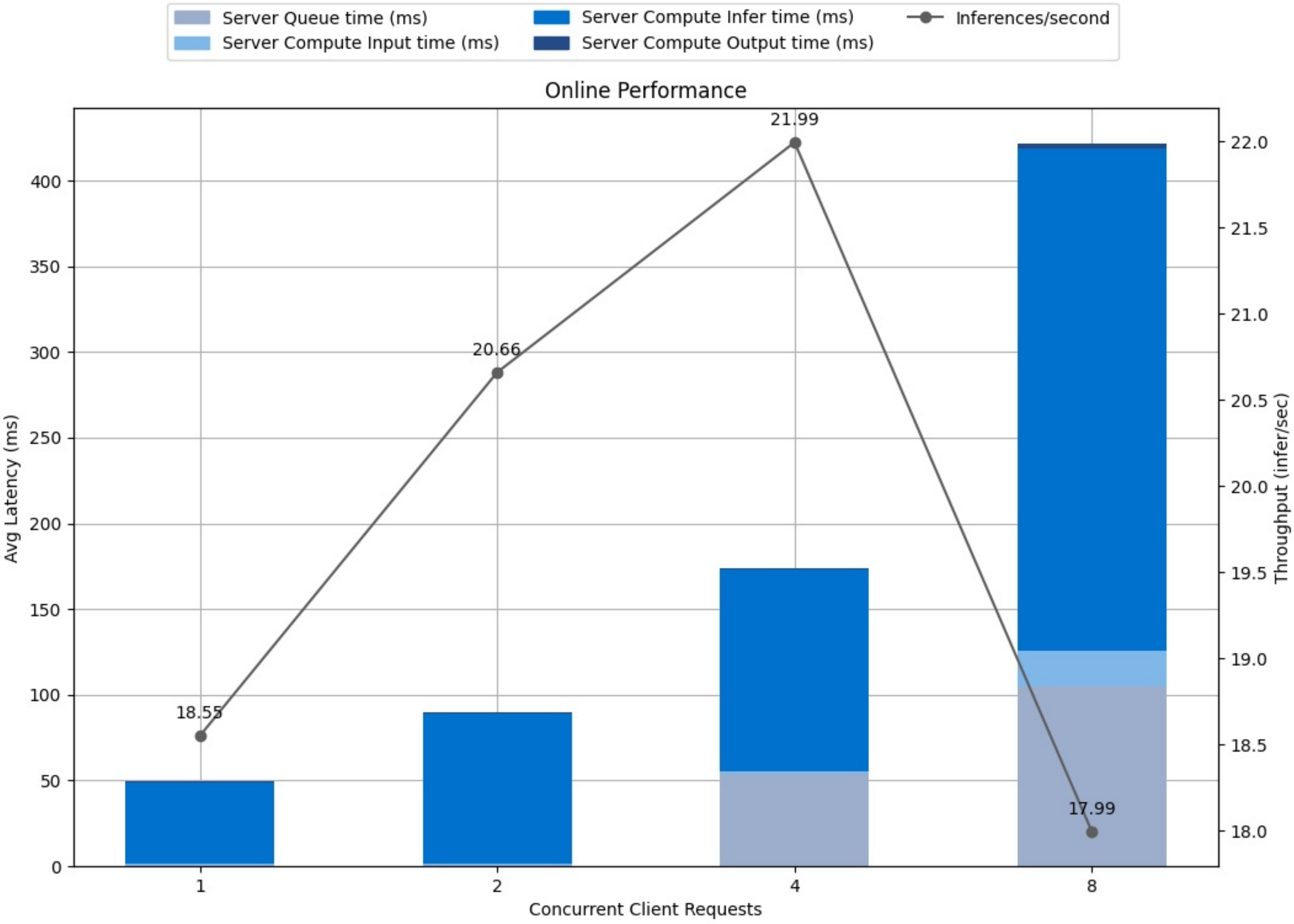
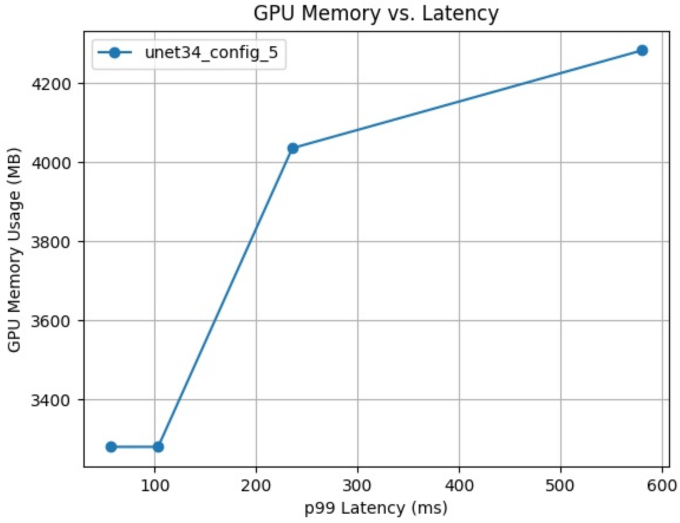
Detailed Report

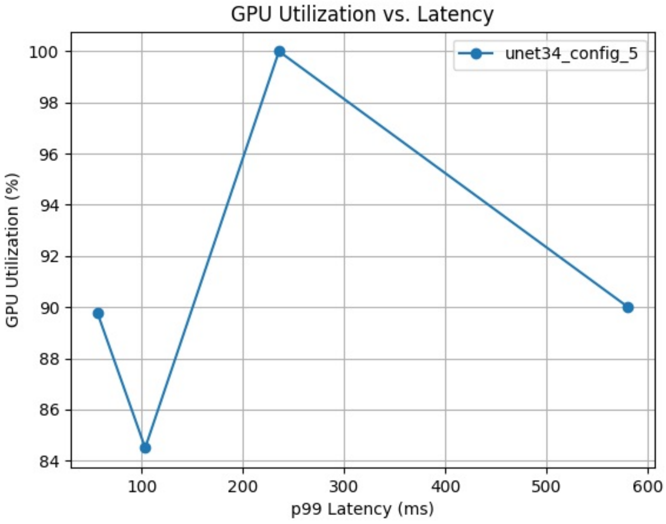
Model Config: unet34_config_5



Latency Breakdown for Online Performance of unet34_config_5

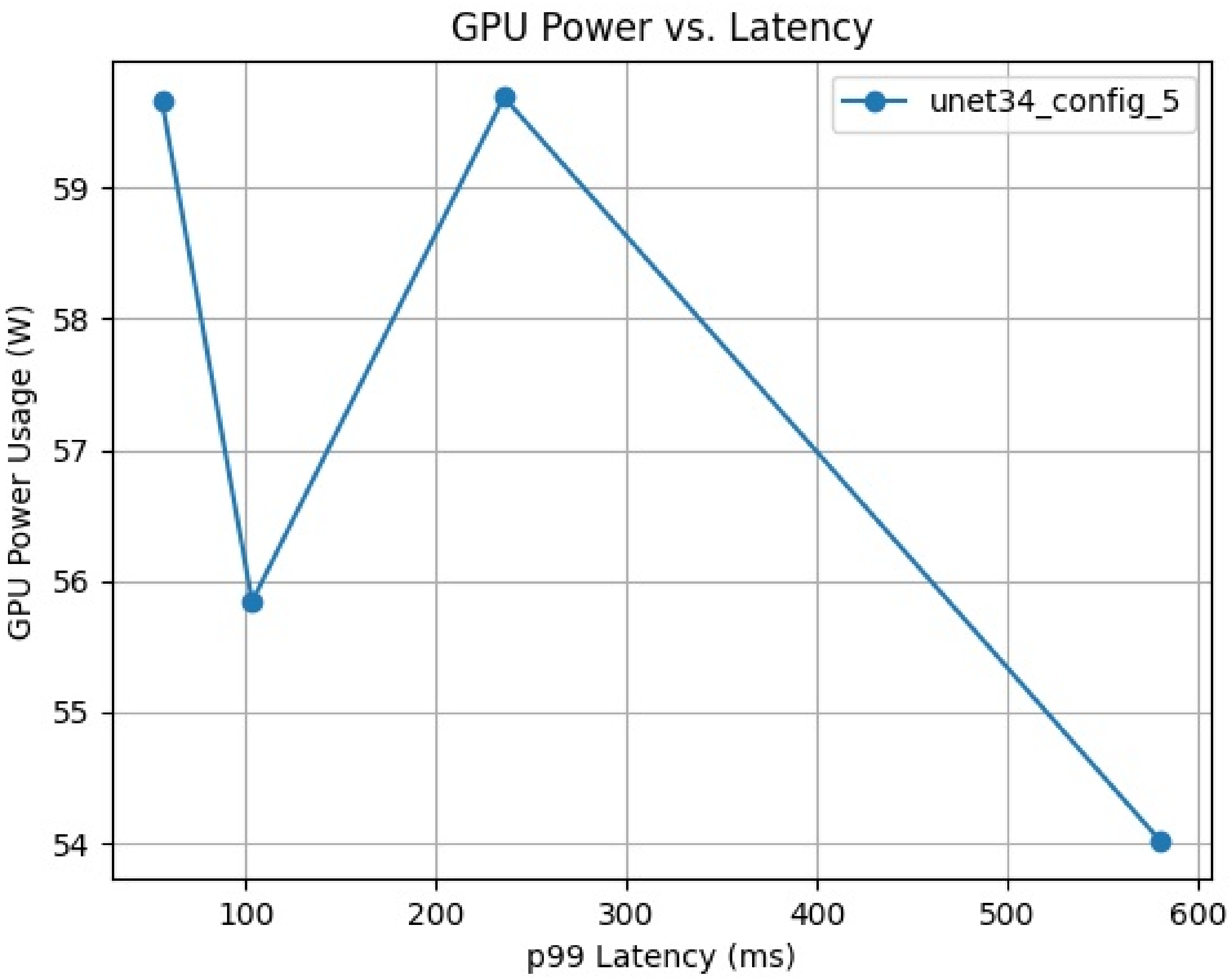


GPU Memory vs. Latency curves for config unet34_config_5



GPU Utilization vs. Latency curves for config unet34_config_5

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 (%)
8	580.471	433.128	105.007	20.998	293.004	17.9946	4282.384384	90.0
4	235.588	181.513	53.539	1.837	117.486	21.9929	4034.920448	100.0
2	103.468	95.348	0.168	1.257	87.702	20.659	3279.945728	84.5
1	56.081	53.138	0.204	1.14	47.665	18.5505	3279.945728	89.8



GPU Power vs. Latency curves for config unet34_config_5

The model config "unet34_config_5" uses 2 GPU instances with a max batch size of 4 and has dynamic batching enabled. 4 measurement(s) were obtained for the model config on GPU(s) 1 x NVIDIA GeForce RTX 3050 Ti Laptop GPU with total memory 3.8 GB. This model uses the platform pytorch_libtorch.

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.