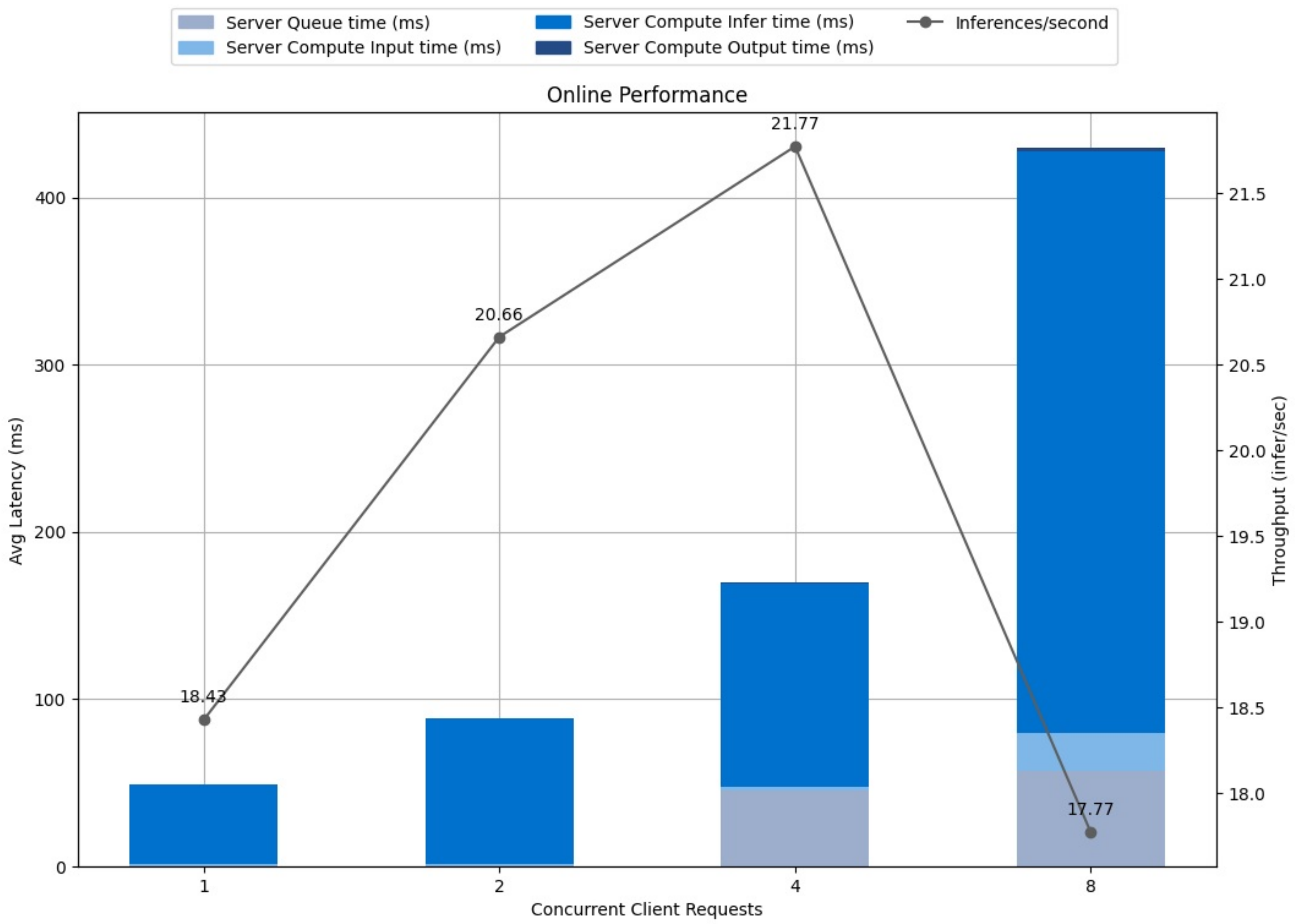
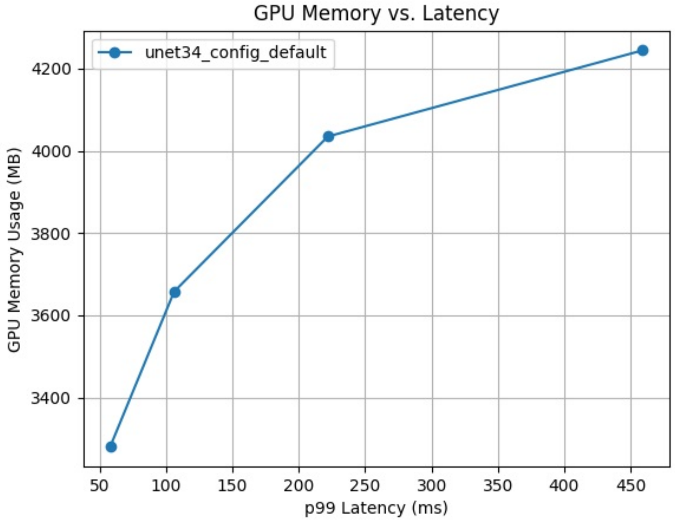
Detailed Report

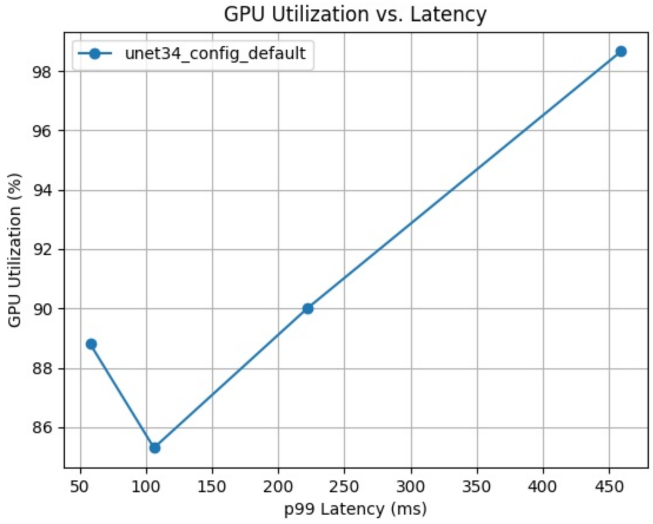
Model Config: unet34_config_default



Latency Breakdown for Online Performance of unet34_config_default

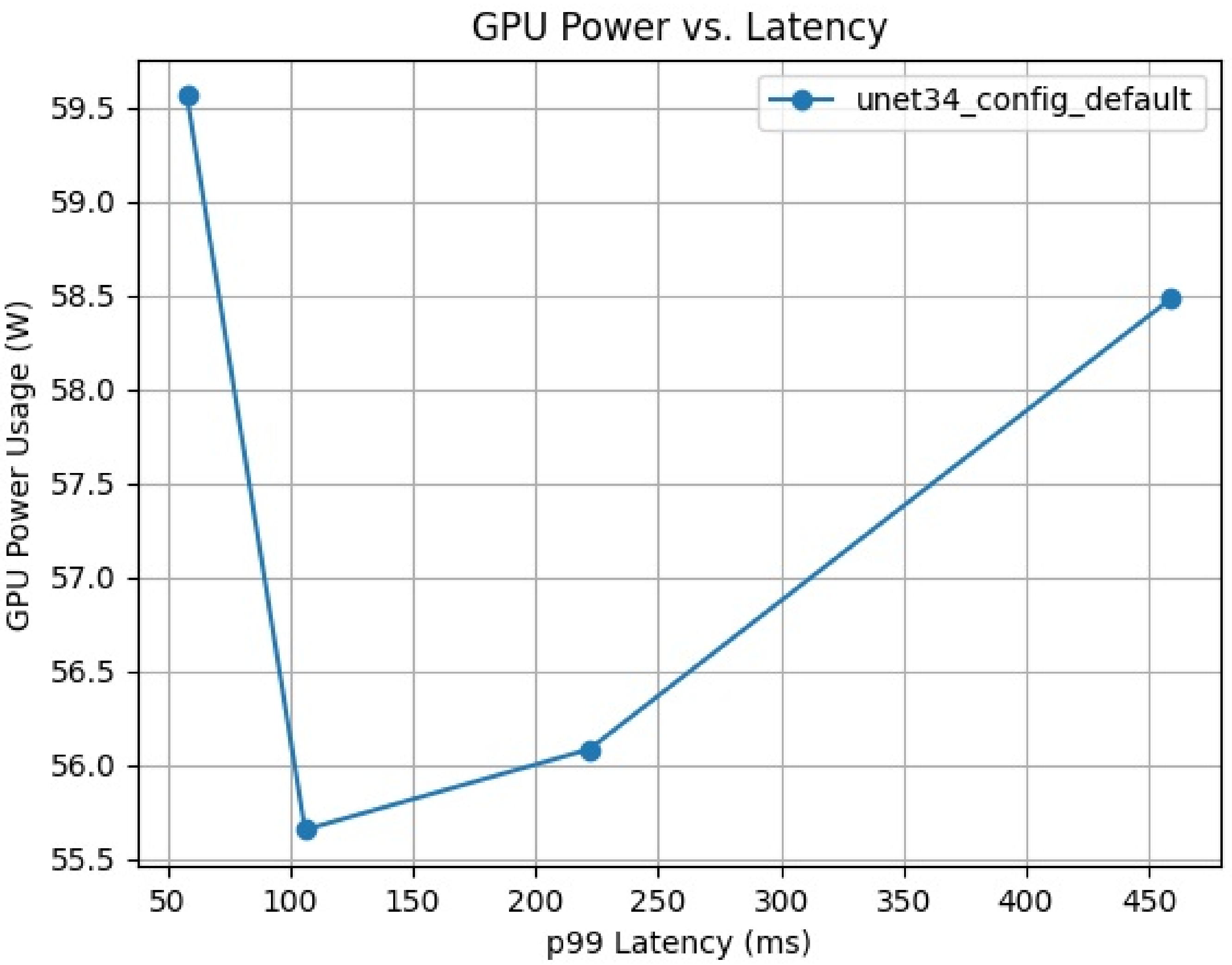


GPU Memory vs. Latency curves for config unet34_config_default



GPU Utilization vs. Latency curves for config unet34_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 GPU Memory Usage (MB)	Average GPU Utilization (%)
8	459.266	444.185	57.411	22.285	347.969	17.7731	4244.635648	98.7
4	221.899	180.056	45.036	2.367	121.694	21.7725	4034.920448	90.0
2	106.0	96.109	0.121	1.465	86.837	20.6618	3657.433088	85.3
1	57.806	53.541	0.147	1.248	47.537	18.4275	3279.945728	88.8



GPU Power vs. Latency curves for config unet34_config_default

The model config "unet34_config_default" uses 2 GPU instances with a max batch size of 6 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.