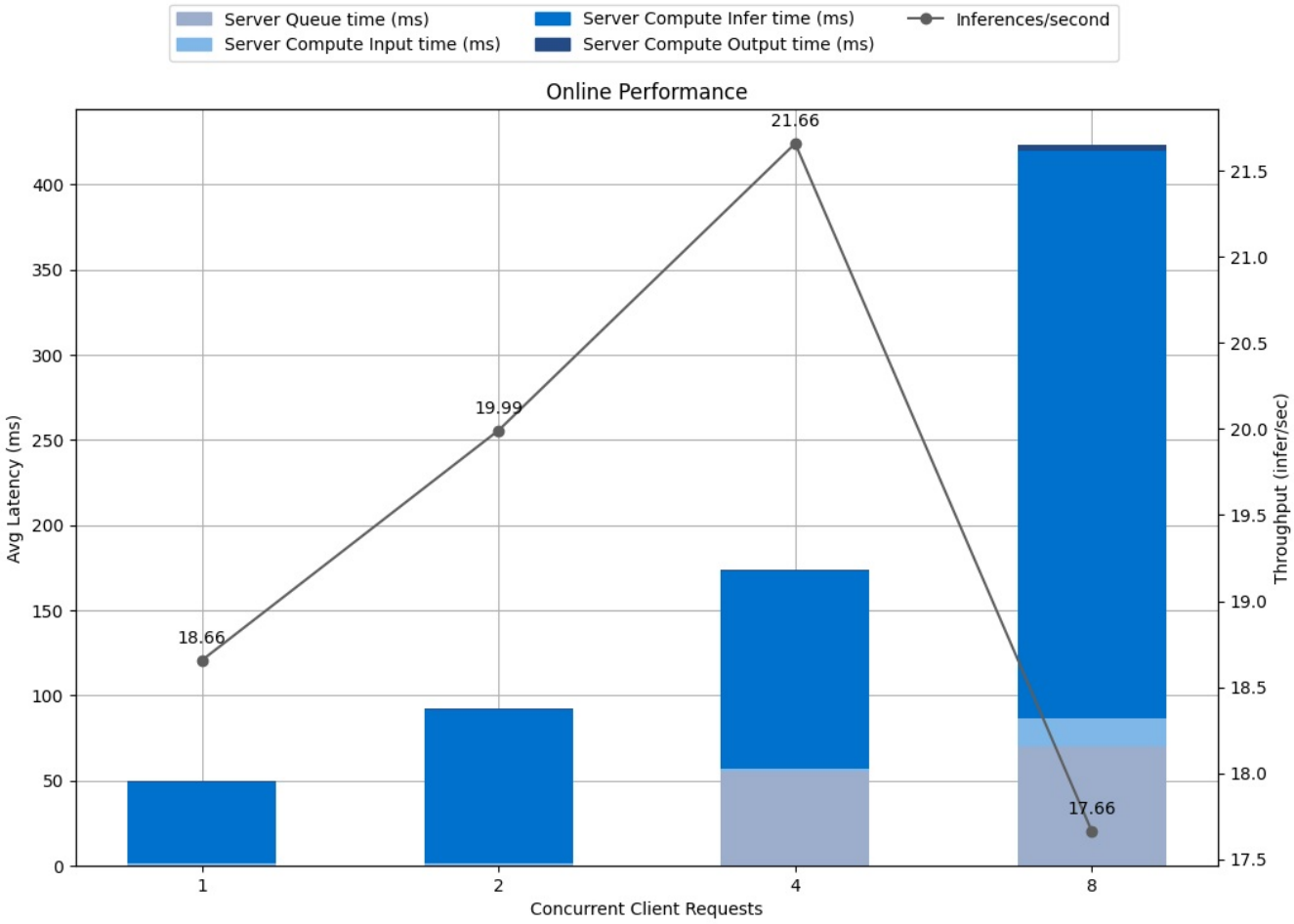
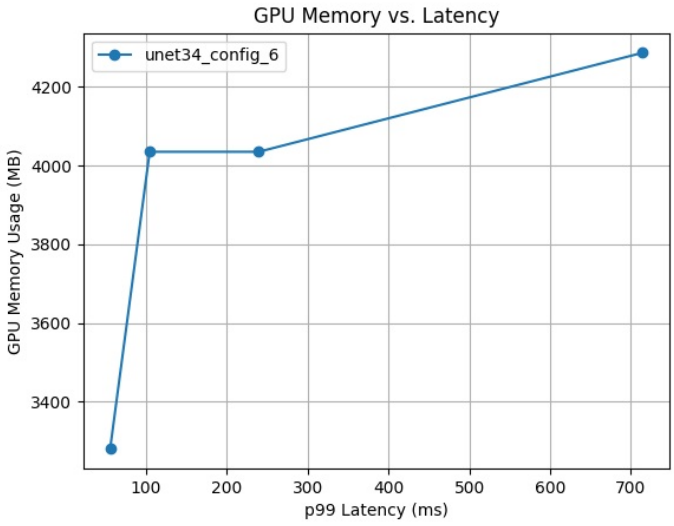
Detailed Report

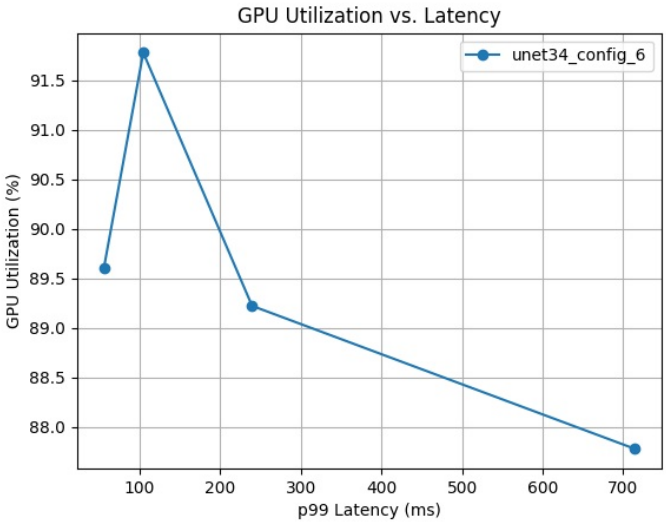
Model Config: unet34_config_6



Latency Breakdown for Online Performance of unet34_config_6

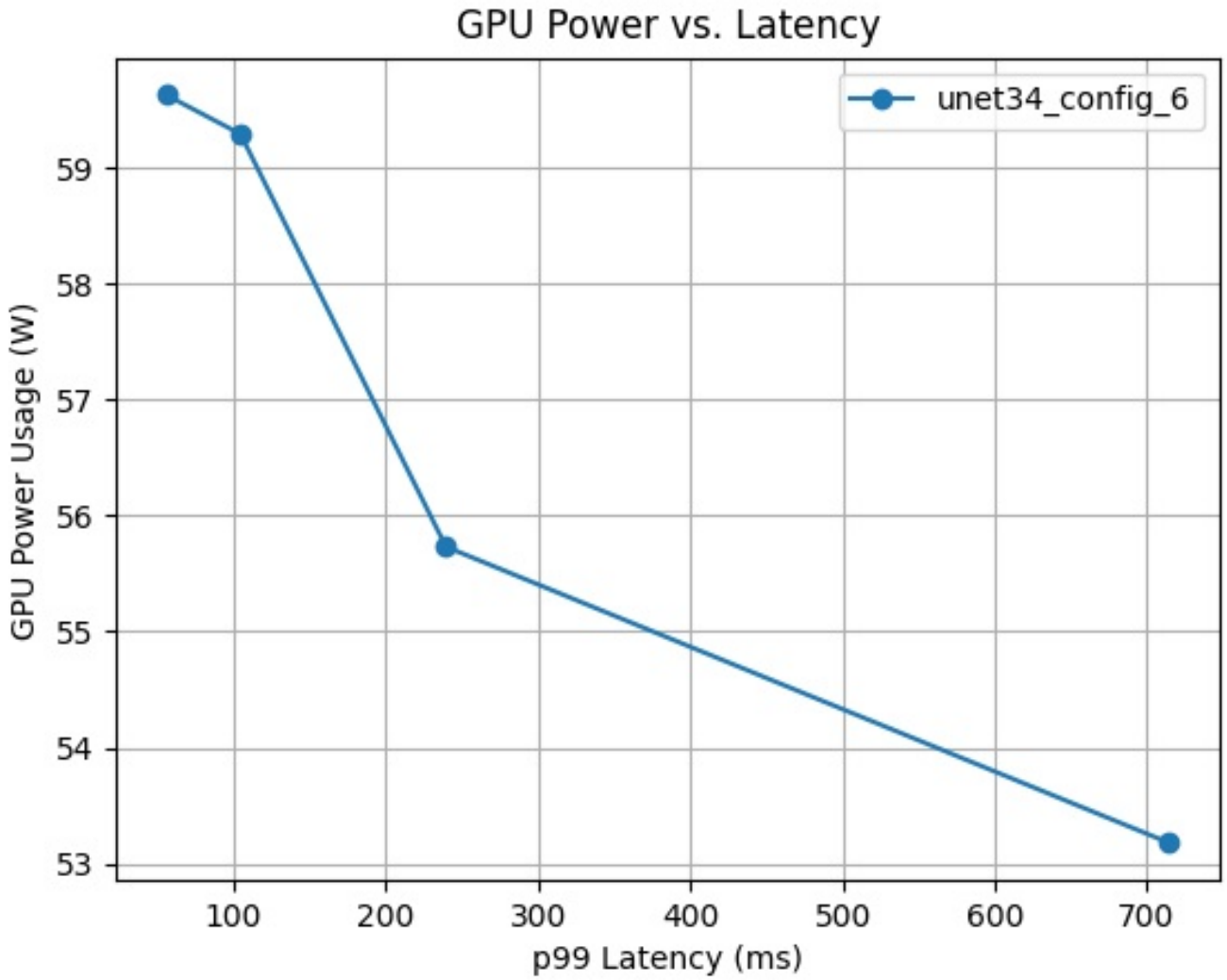


GPU Memory vs. Latency curves for config unet34_config_6



GPU Utilization vs. Latency curves for config unet34_config_6

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	714.847	436.113	70.322	16.204	333.214	17.6613	4286.578688	87.8
4	239.376	181.492	55.383	1.85	115.879	21.6599	4034.920448	89.2
2	104.268	98.854	0.151	1.407	90.094	19.9924	4034.920448	91.8
1	55.281	53.083	0.202	1.139	47.706	18.6557	3279.945728	89.6



GPU Power vs. Latency curves for config unet34_config_6

The model config "unet34_config_6" uses 2 GPU instances with a max batch size of 8 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.