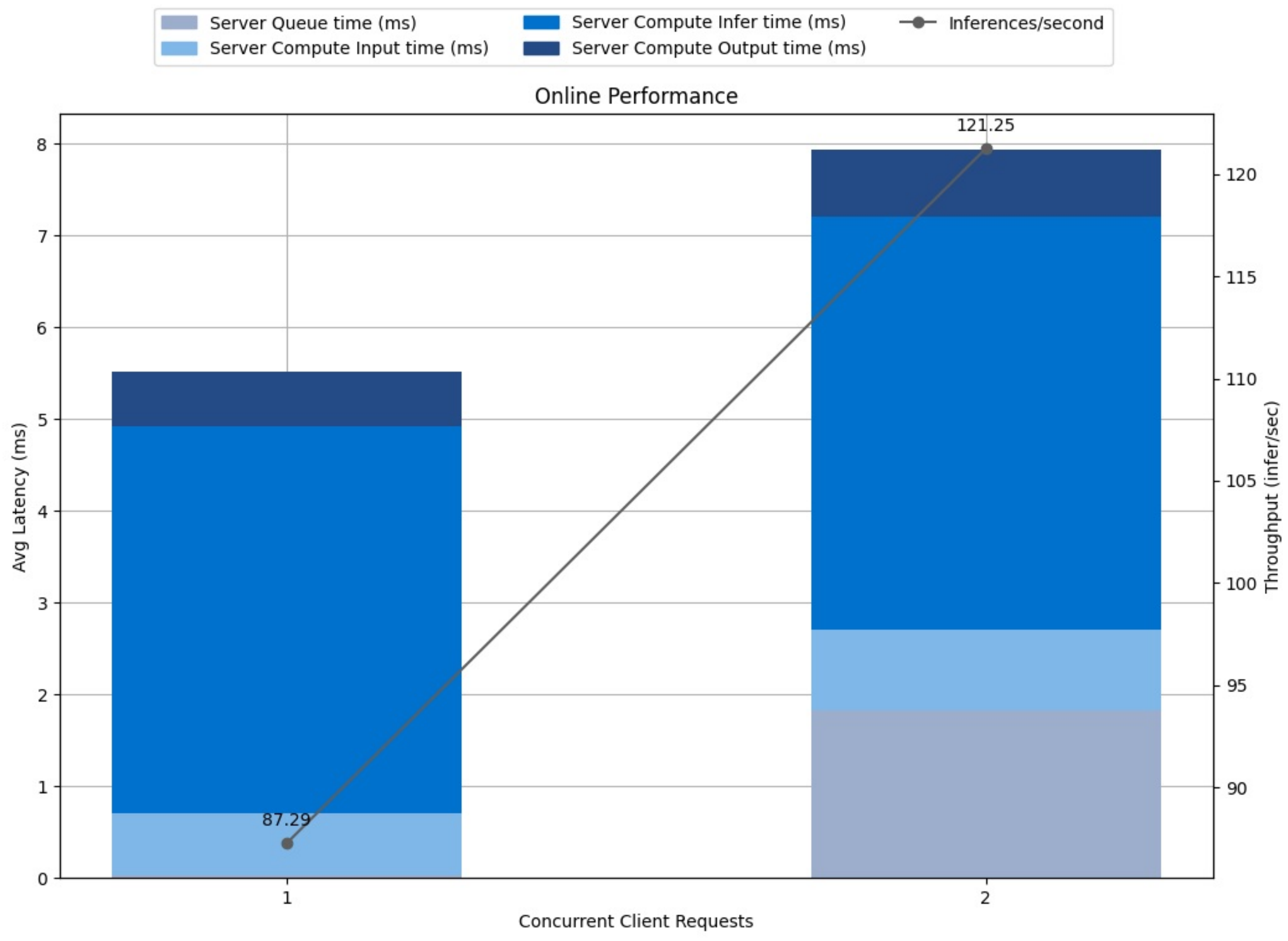
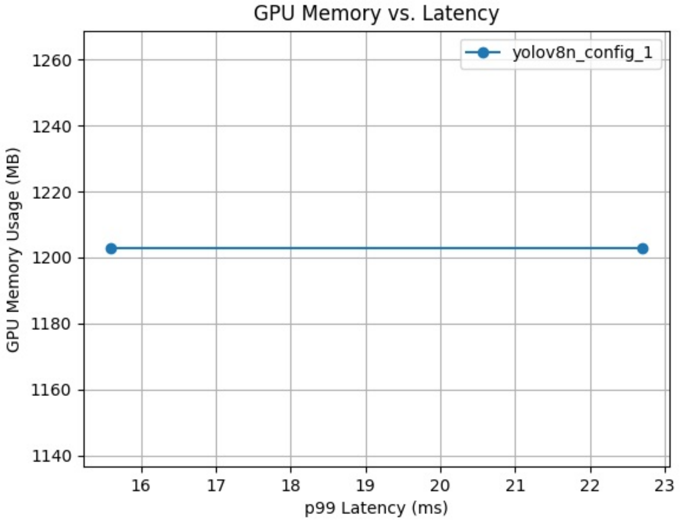
Detailed Report

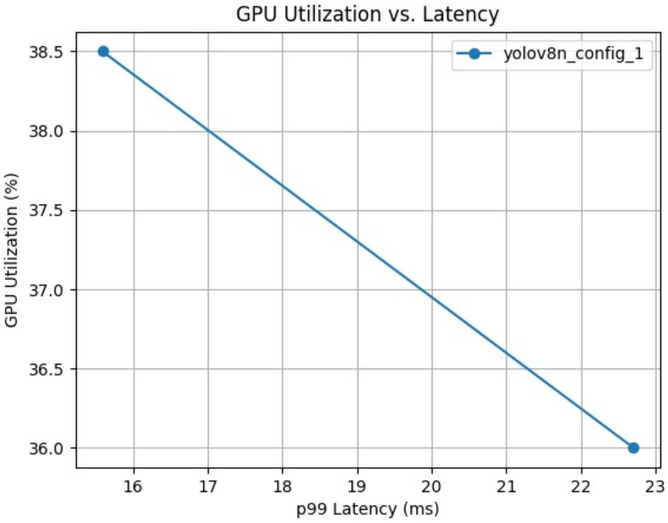
Model Config: yolov8n_config_1



Latency Breakdown for Online Performance of yolov8n_config_1

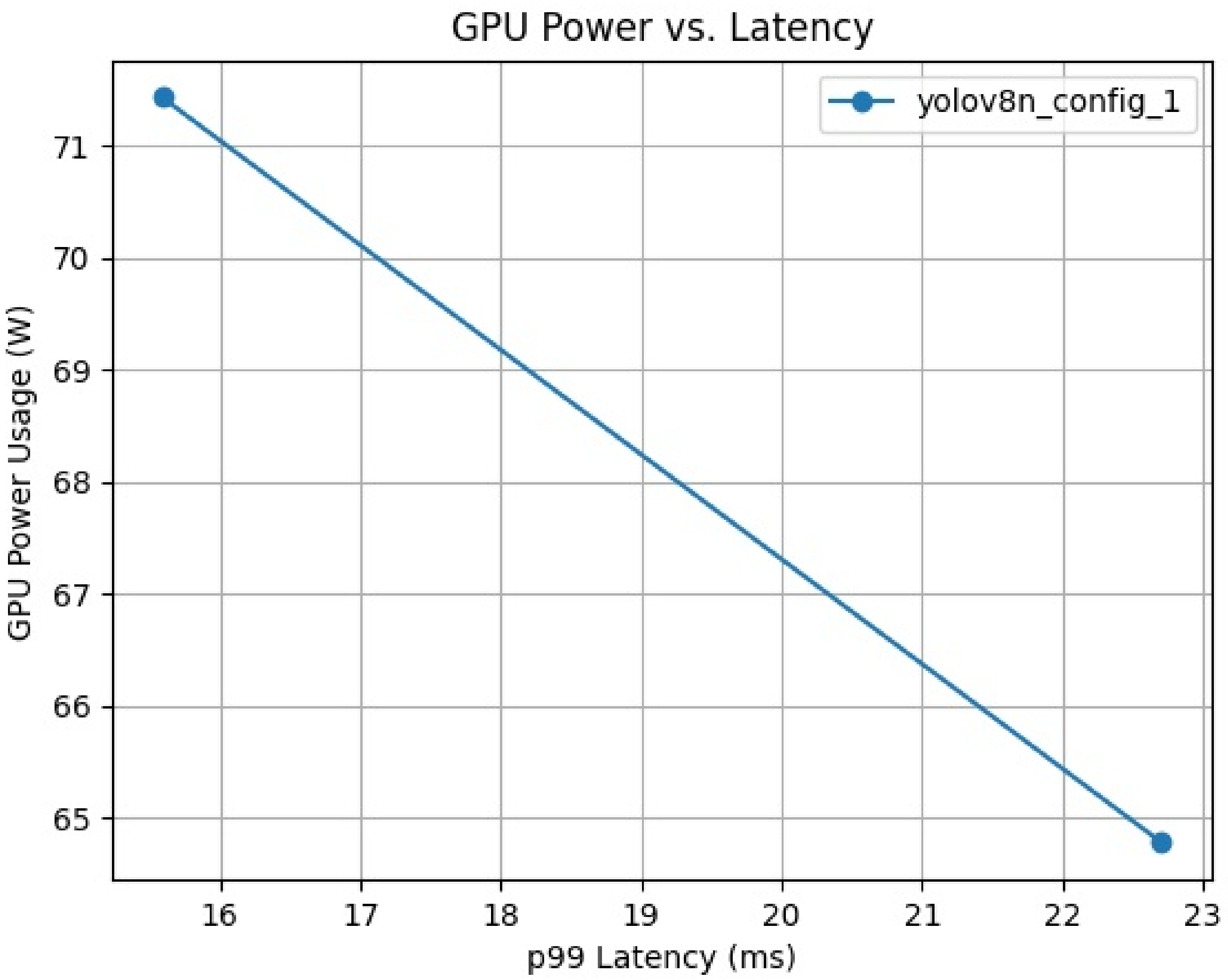


GPU Memory vs. Latency curves for config yolov8n_config_1



GPU Utilization vs. Latency curves for config yolov8n_config_1

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 (%)
2	22.706	15.837	1.819	0.89	4.493	121.254	1202.716672	36.0
1	15.589	10.848	0.029	0.67	4.226	87.2887	1202.716672	38.5



GPU Power vs. Latency curves for config yolov8n_config_1

The model config **yolov8n_config_1** uses 1 GPU instance with a max batch size of 2 and has dynamic batching enabled. 2 measurement(s) were obtained for the model config on GPU(s) 1 x NVIDIA GeForce RTX 3060 with total memory 11.7 GB. This model uses the platform .

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 latency.