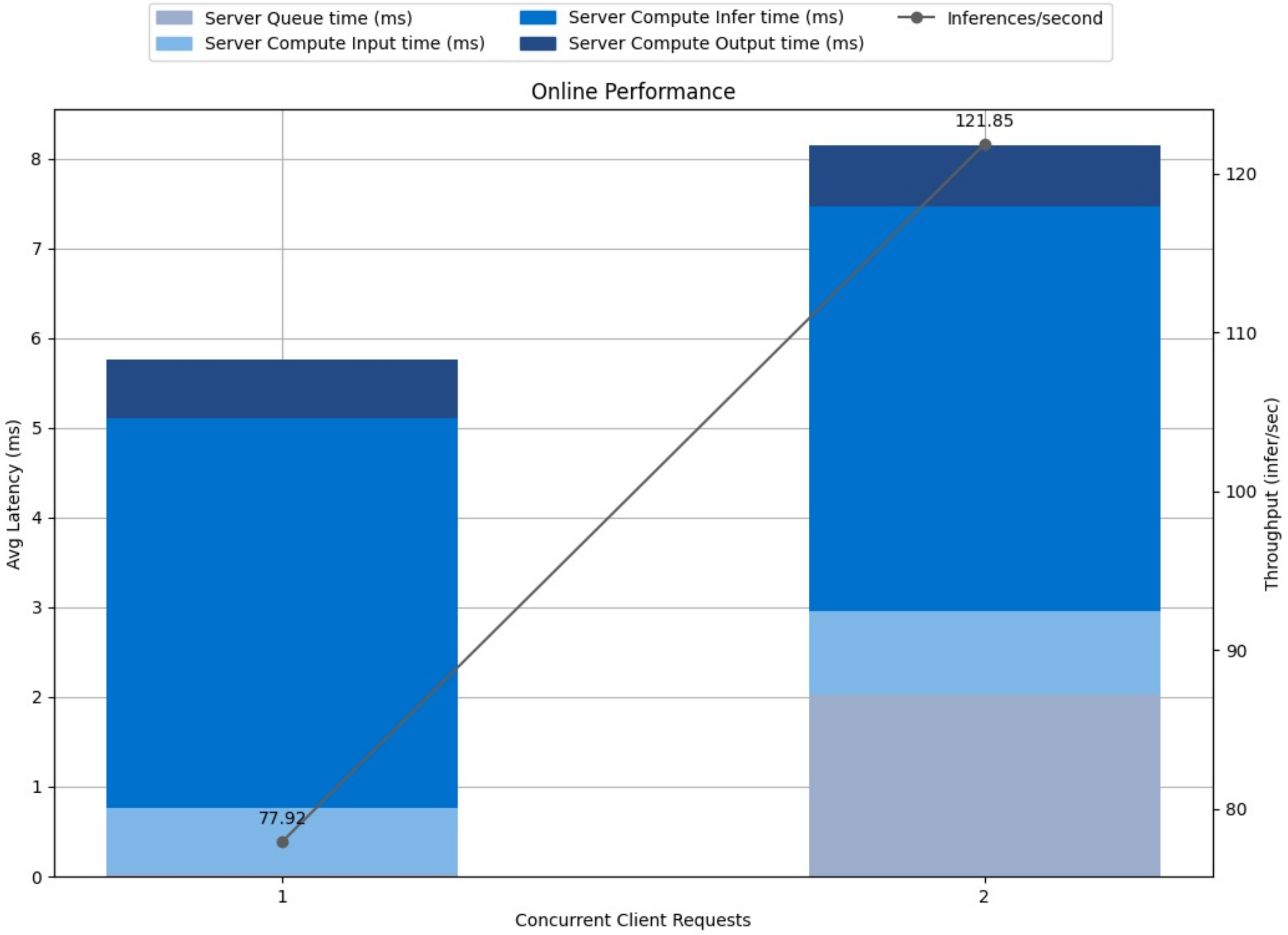
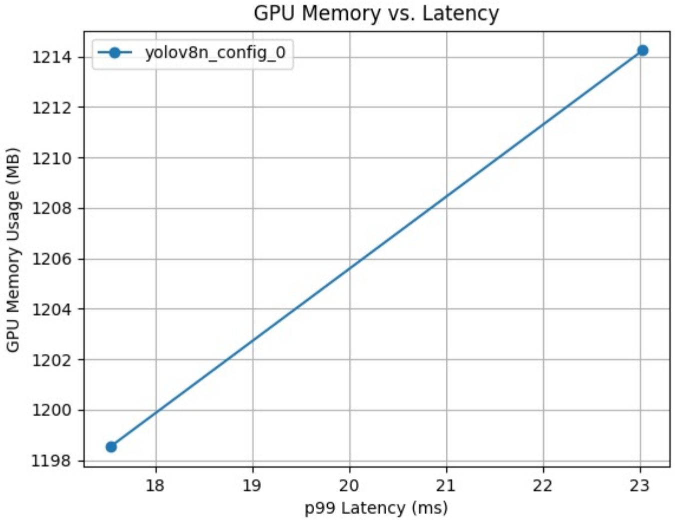
Detailed Report

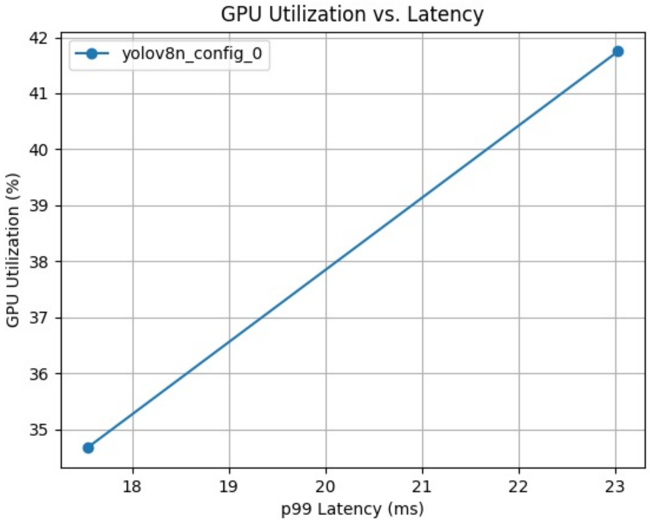
Model Config: yolov8n_config_0



Latency Breakdown for Online Performance of yolov8n_config_0

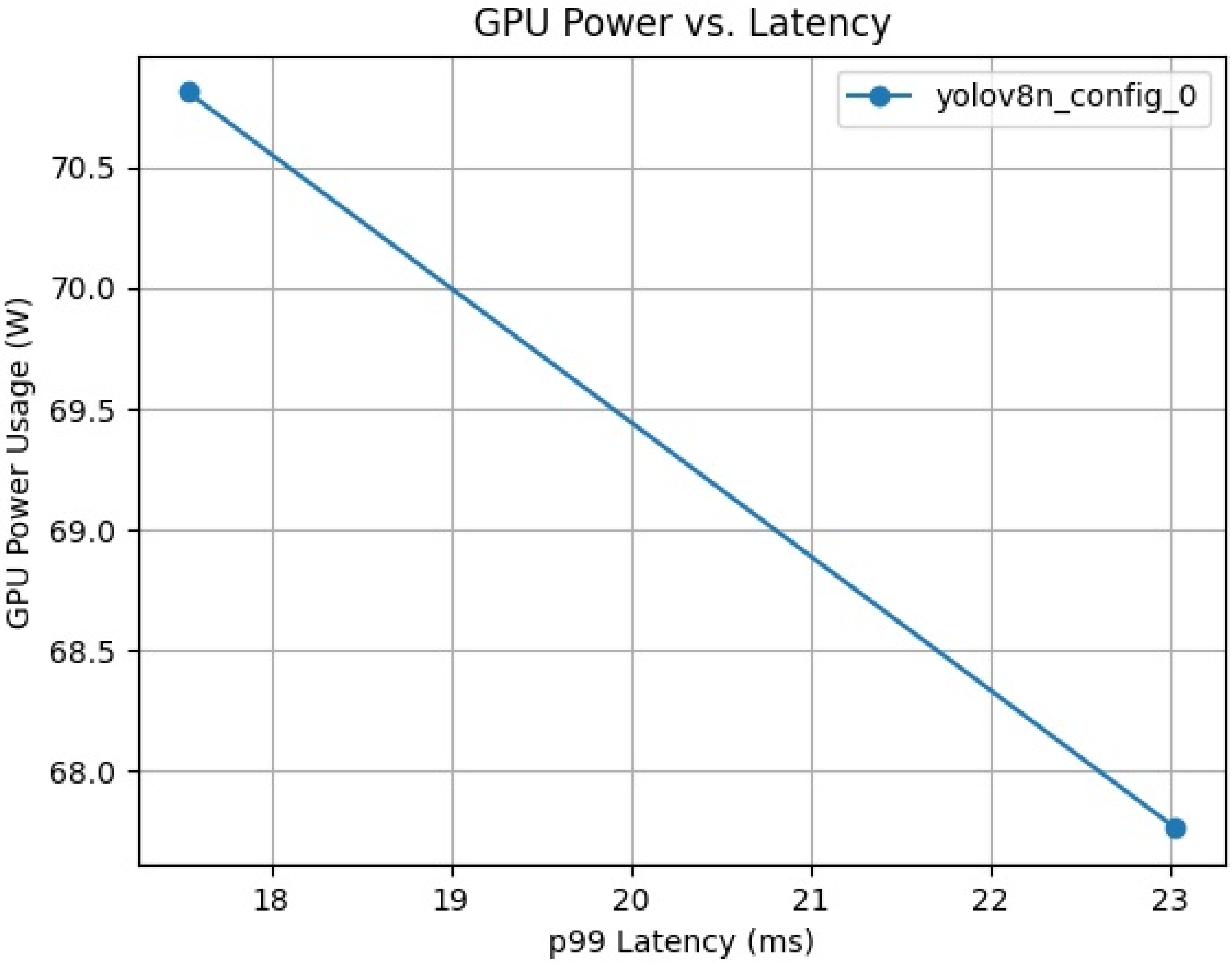


GPU Memory vs. Latency curves for config yolov8n_config_0



GPU Utilization vs. Latency curves for config yolov8n_config_0

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	23.035	15.728	2.026	0.934	4.505	121.852	1214.251008	41.8
1	17.535	12.254	0.032	0.735	4.341	77.9224	1198.522368	34.7



GPU Power vs. Latency curves for config yolov8n_config_0

The model config **yolov8n_config_0** uses 1 GPU instance with a max batch size of 1 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.