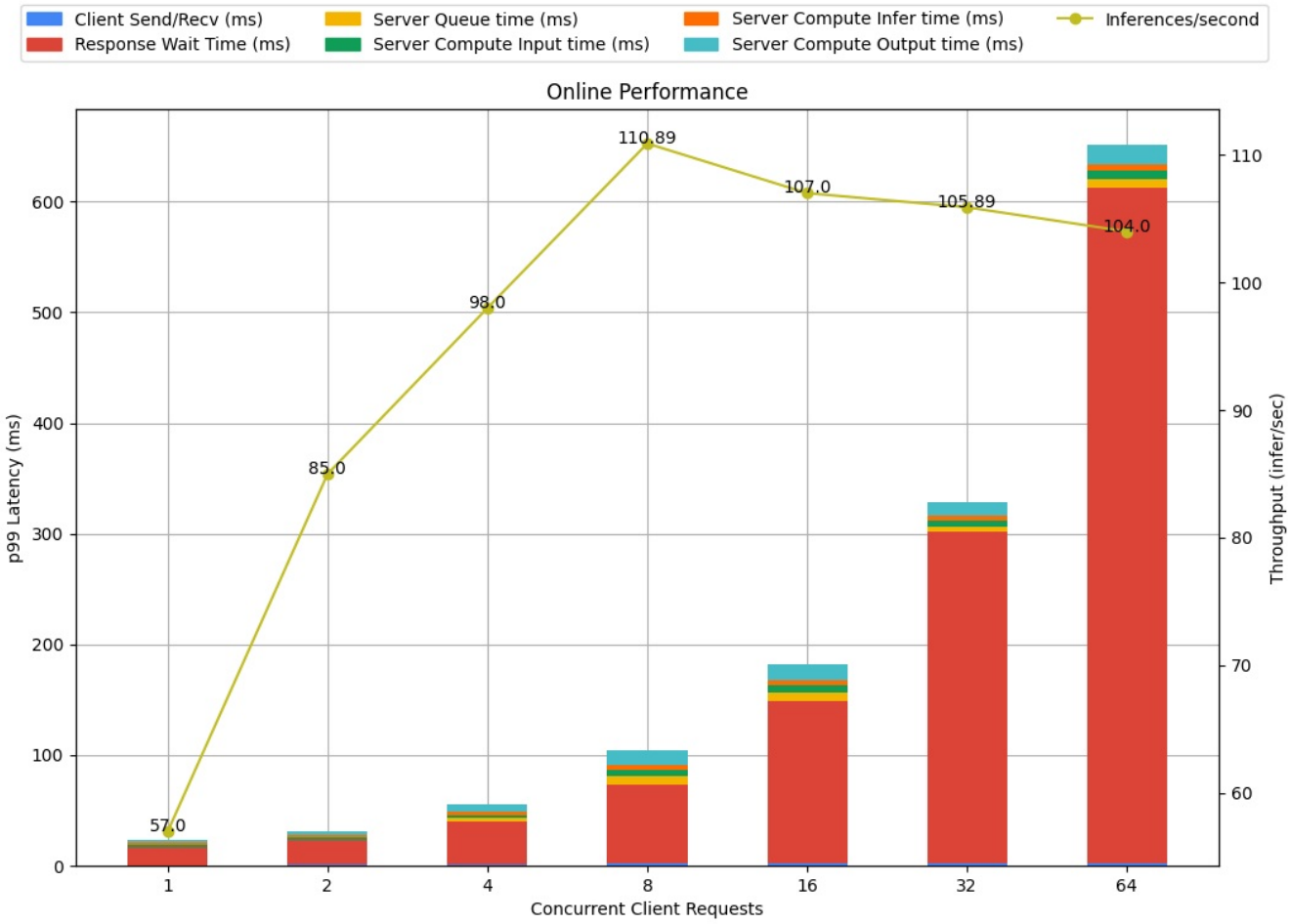
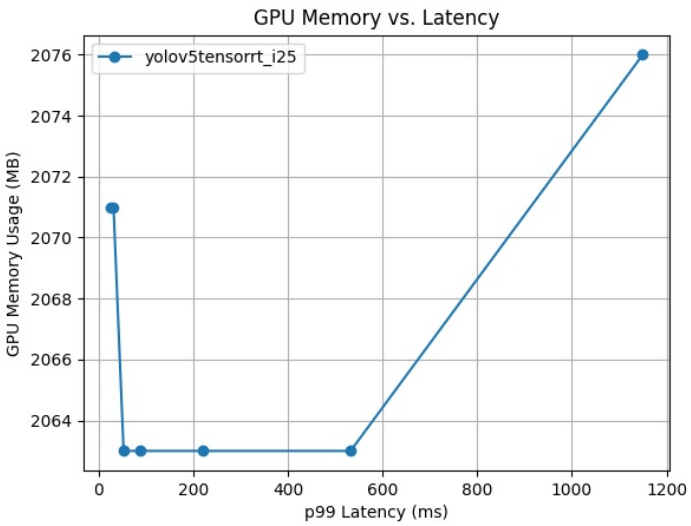
Detailed Report

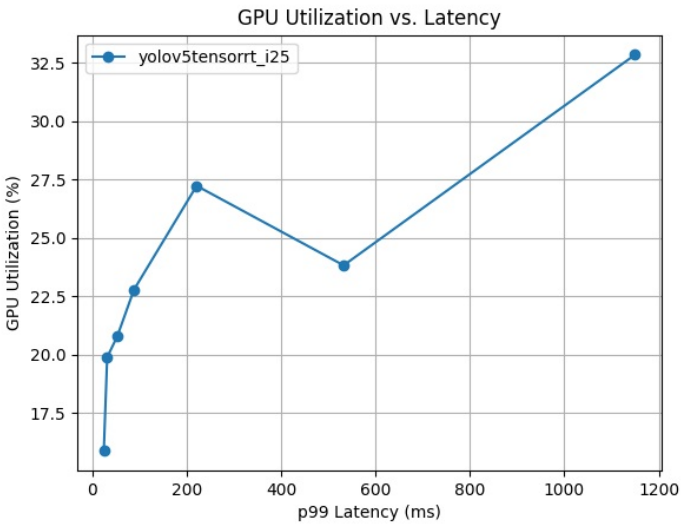
Model Config: yolov5tensorrt_i25



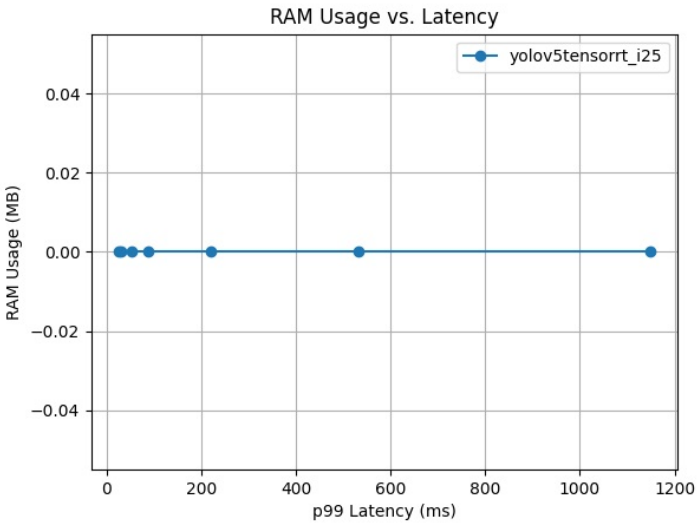
Latency Breakdown for Online Performance of yolov5tensorrt_i25



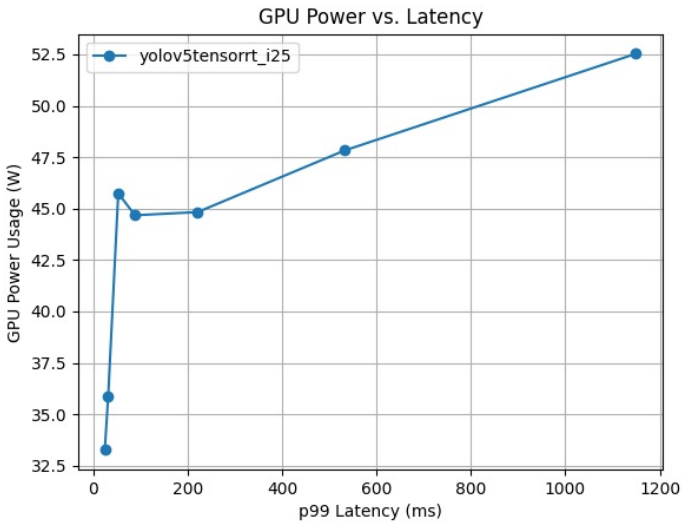
GPU Memory vs. Latency curves for config yolov5tensorrt_i25



GPU Utilization vs. Latency curves for config yolov5tensorrt_i25



RAM Usage vs. Latency curves for config yolov5tensorrt_i25



GPU Power vs. Latency curves for config yolov5tensorrt_i25

Request Concurrency	p99 Latency (ms)	Client Response Wait (ms)	Server Queue (ms)	Server Compute Input (ms)	Server Compute Infer (ms)	Throughput (infer/sec)	Max CPU Memory Usage (MB)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
64	1149.712	611.28	7.782	7.45	5.811	104.0	0	2076.0	32.8
32	532.459	299.854	4.959	5.44	4.705	105.894	0	2063.0	23.8
16	221.147	146.964	7.218	6.606	5.334	107.0	0	2063.0	27.2
8	88.393	71.365	7.542	6.082	4.465	110.889	0	2063.0	22.8
4	52.969	38.235	3.106	3.044	3.503	98.0	0	2063.0	20.8
2	31.648	22.398	0.284	1.87	2.789	85.0	0	2071.0	19.9
1	24.474	16.434	0.056	1.514	2.657	57.0	0	2071.0	15.9

The model config "yolov5tensorrt_i25" uses 1 GPU instances. 7 measurements were obtained for the model config on GPU(s) NVIDIA GeForce RTX 3070 Laptop GPU with memory limit(s) 7.8 GB. This model uses the platform tensorrt_plan. This model config has dynamic batching enabled with preferred batch size(s) [8].

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.