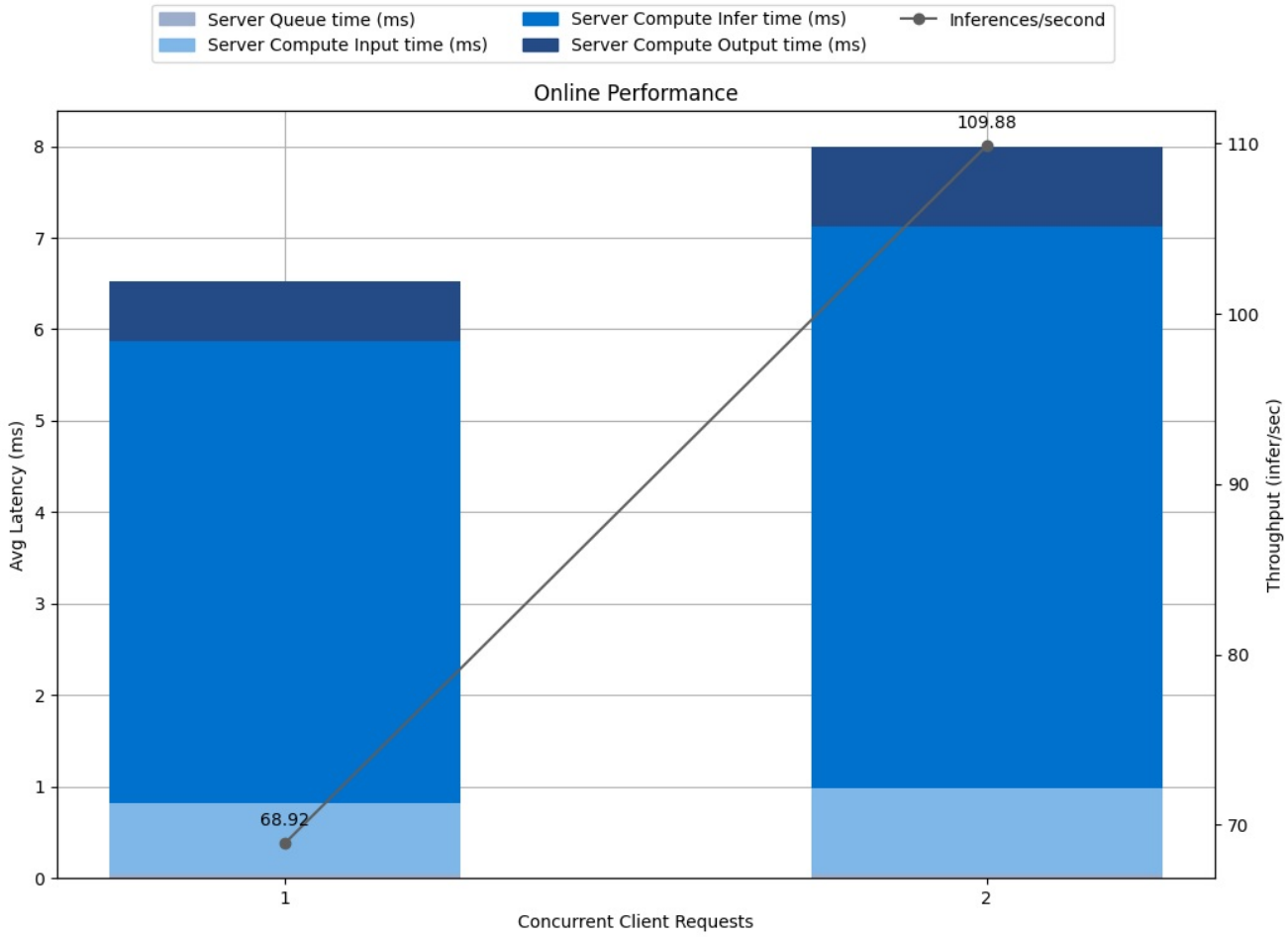
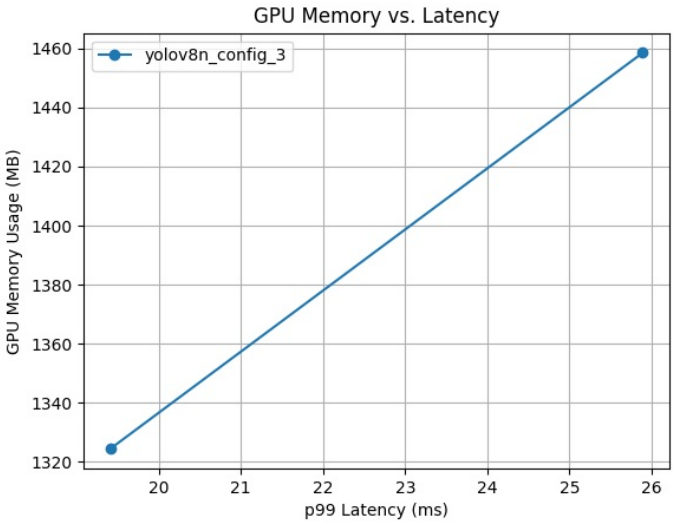
# Detailed Report

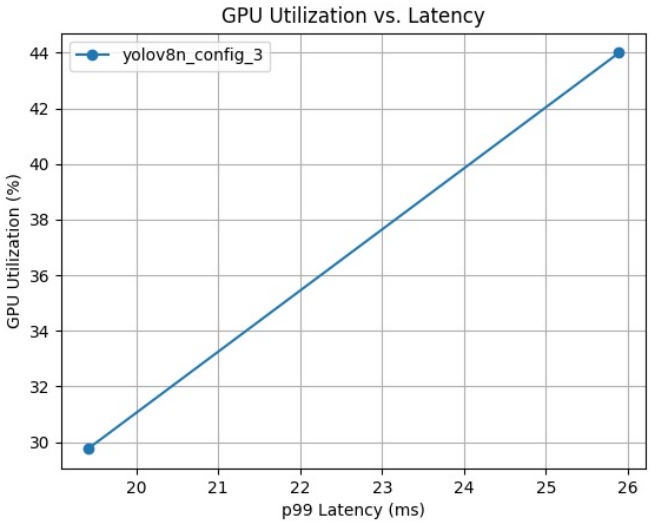
## Model Config: yolov8n_config_3




**Latency Breakdown for Online Performance of yolov8n_config_3**

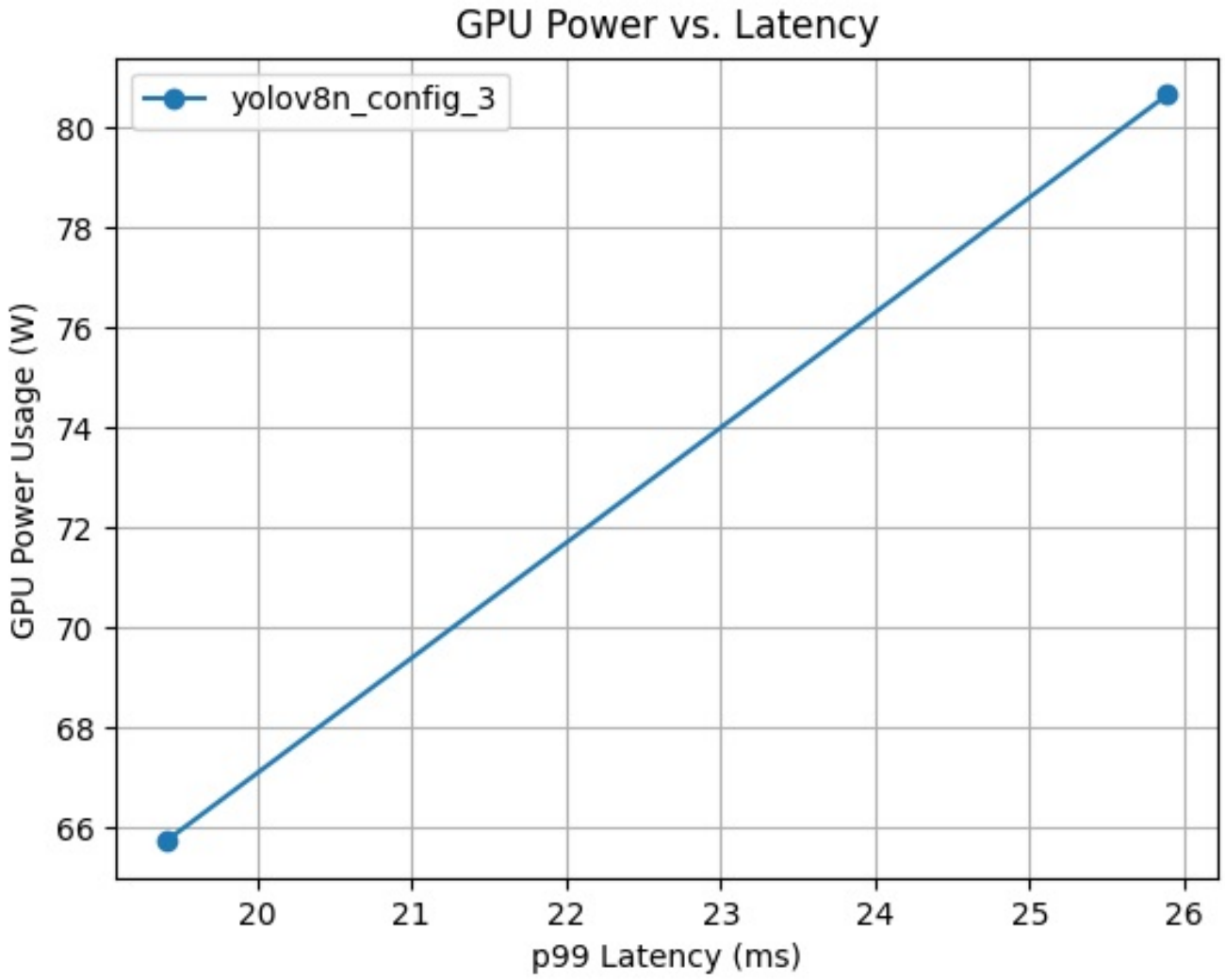



**GPU Memory vs. Latency curves for config yolov8n_config_3**




**GPU Utilization vs. Latency curves for config yolov8n_config_3**

| 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 | 25.891 | 17.343 | 0.04 | 0.935 | 6.148 | 109.879 | 1458.569216 | 44.0 |
| 1 | 19.411 | 13.891 | 0.045 | 0.767 | 5.063 | 68.9228 | 1324.351488 | 29.7 |




**GPU Power vs. Latency curves for config yolov8n_config_3**

The model config **yolov8n_config_3** uses 2 GPU instances 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.