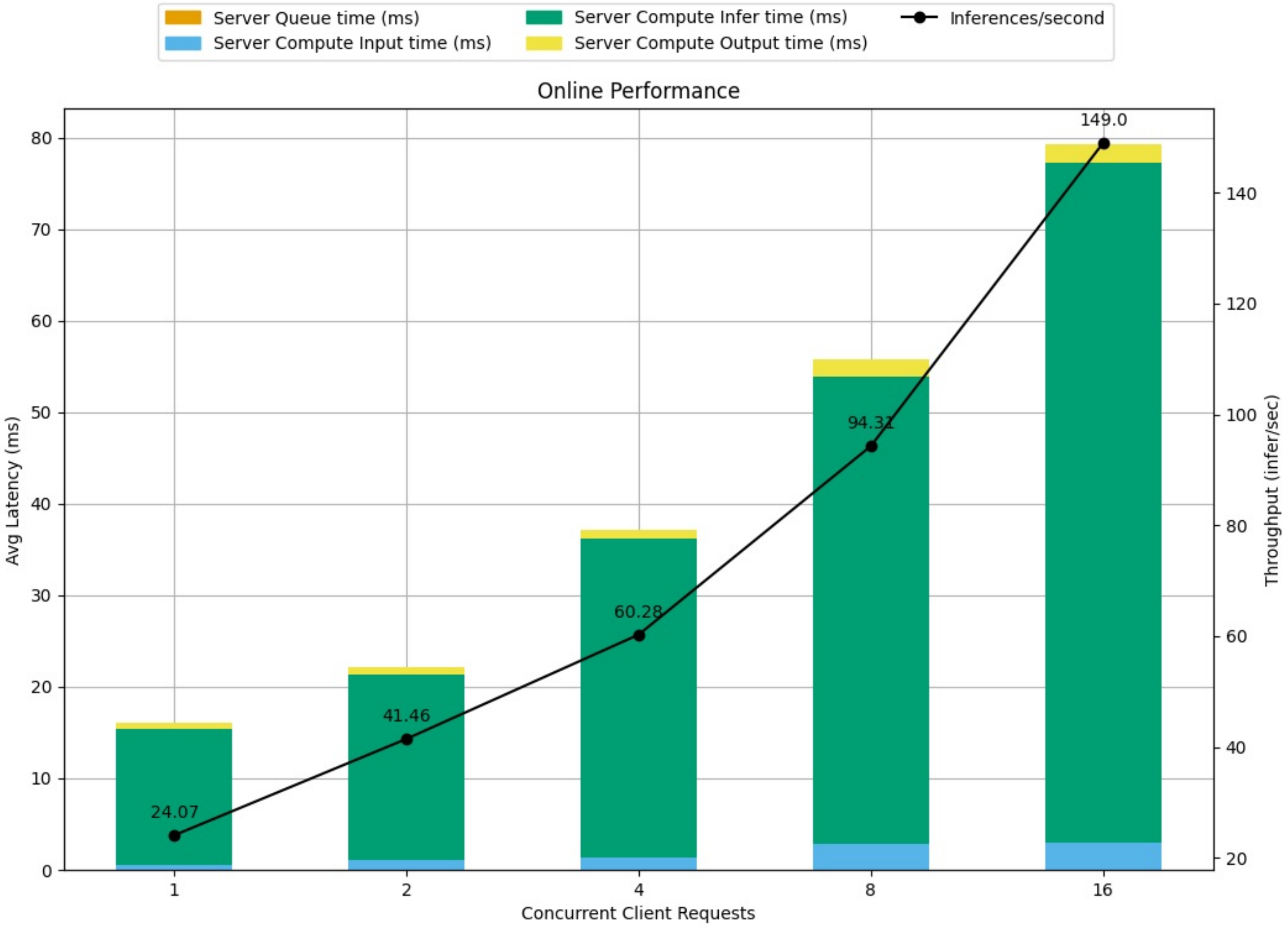
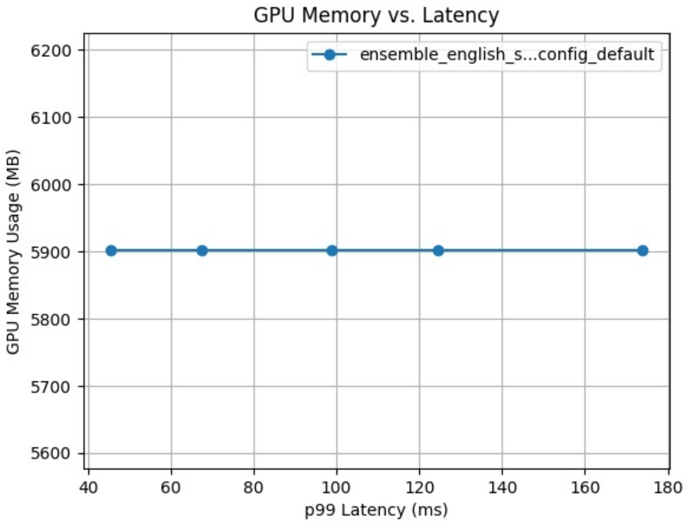
# Detailed Report

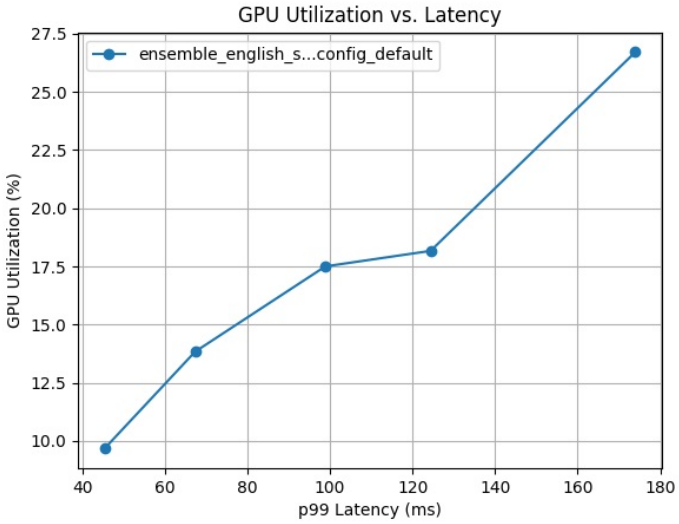
## Model Config: ensemble_english_stt_tensorrt_config_default



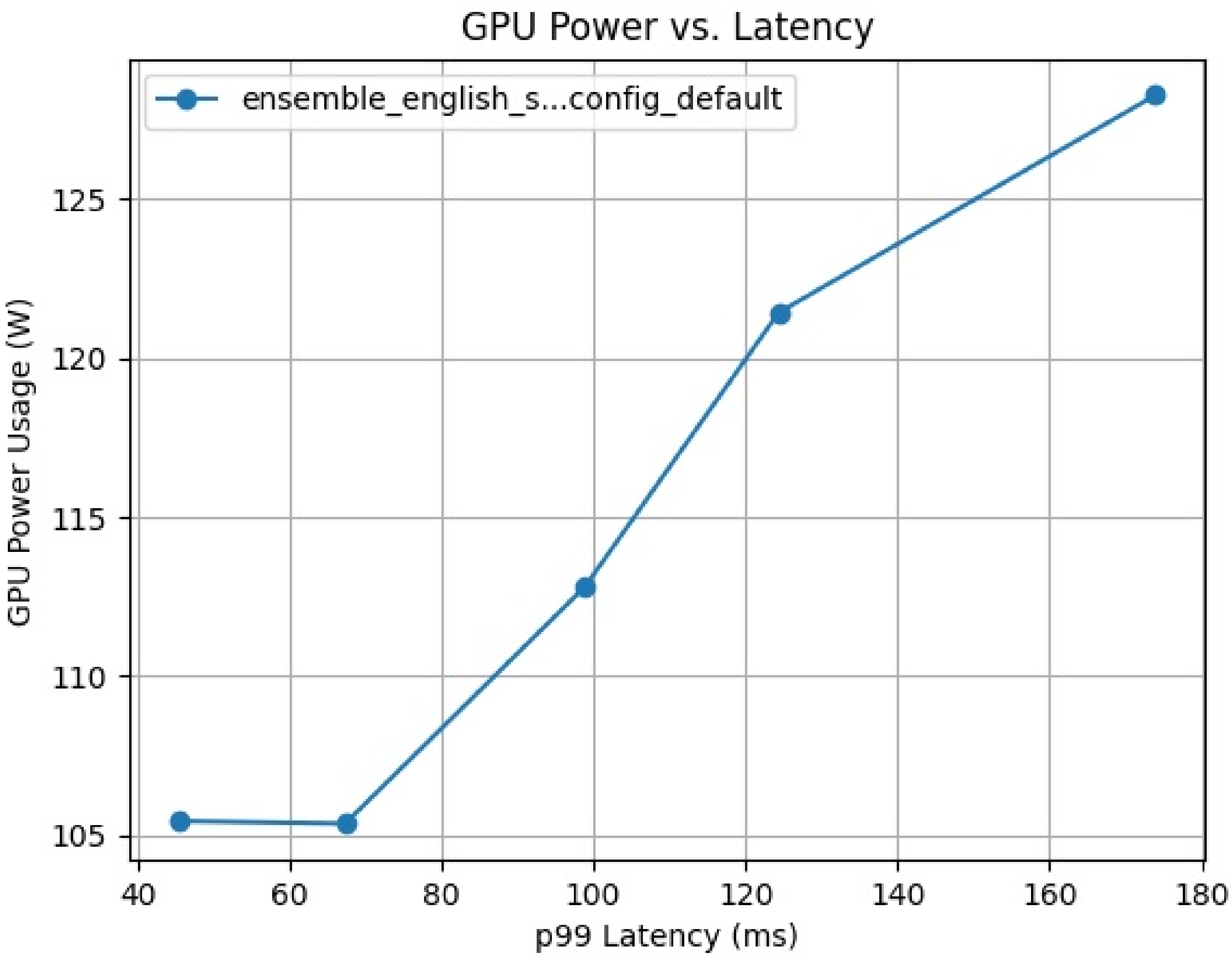
**Latency Breakdown for Online Performance of ensemble_english_stt_tensorrt_config_default**



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



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



GPU Power vs. Latency curves for config ensemble_english_stt_tensorrt_config_default

| 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 (%) |
|---|---|---|---|---|---|---|---|---|
| 16 | 173.979 | 103.56 | 0.002 | 3.035 | 74.197 | 148.999 | 5901.385728 | 26.7 |
| 8 | 124.373 | 81.958 | 0.003 | 2.829 | 51.109 | 94.3082 | 5901.385728 | 18.2 |
| 4 | 98.877 | 64.119 | 0.003 | 1.321 | 34.885 | 60.279 | 5901.385728 | 17.5 |
| 2 | 67.331 | 45.385 | 0.003 | 1.13 | 20.207 | 41.4578 | 5901.385728 | 13.8 |
| 1 | 45.369 | 38.461 | 0.004 | 0.591 | 14.826 | 24.0711 | 5901.385728 | 9.7 |

**ensemble_english_stt_tensorrt_config_default** is comprised of the following composing models:

- **preprocessing_english_stt_config_default**: 5 CPU instances with a max batch size of 8 on platform python
- **tensorrt_english_stt_config_default**: 3 GPU instances with a max batch size of 8 on platform tensorrt
- **postprocessing_english_stt_config_default**: 5 CPU instances with a max batch size of 8 on platform python

5 measurement(s) were obtained for the model config on GPU(s) 1 x NVIDIA L40S with total memory 44.4 GB.

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.