# Online Result Summary

## Model: resnet50

GPU(s): 1 x NVIDIA GeForce RTX 3060 Laptop GPU

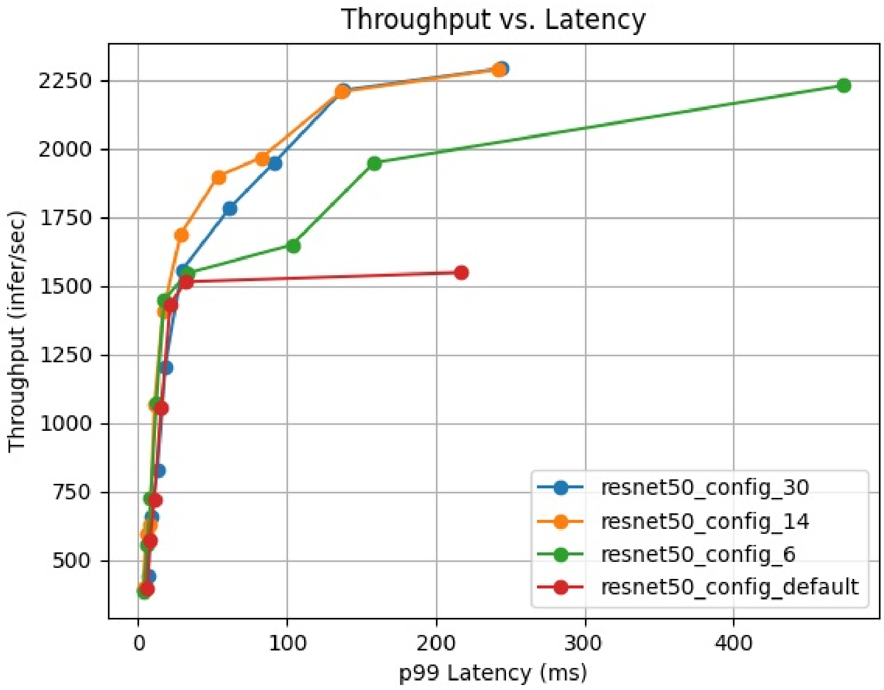
Total Available GPU Memory: 6.0 GB

Constraint targets: None

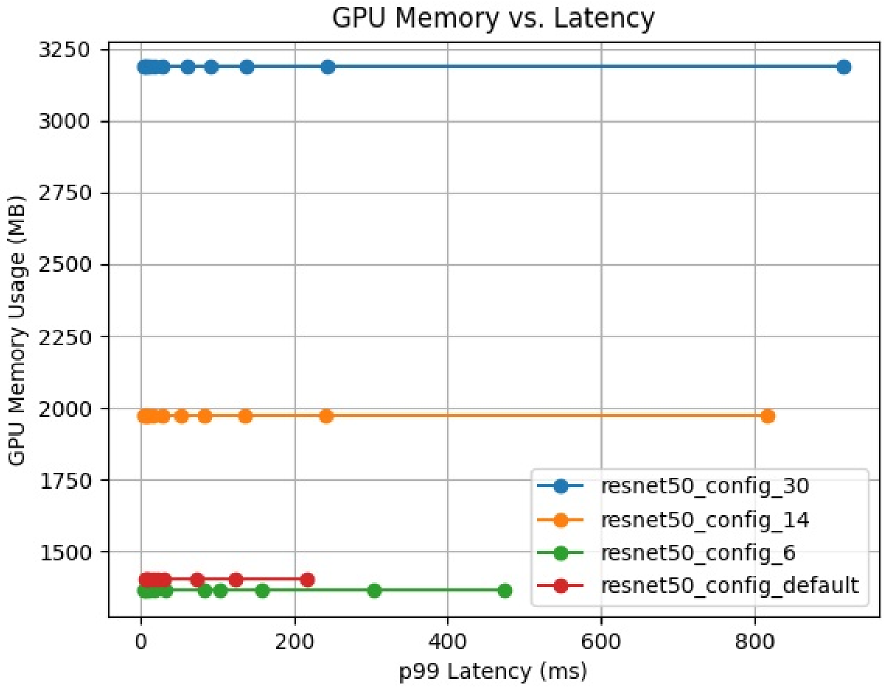
In 378 measurements across 40 configurations, **resnet50_config_30** is **48%** better than the default configuration at maximizing throughput, under the given constraints, on GPU(s) 1 x NVIDIA GeForce RTX 3060 Laptop GPU.

- **resnet50_config_30**: 4 GPU instances with a max batch size of 64 on platform tensorrt_plan

Curves corresponding to the 3 best model configuration(s) out of a total of 40 are shown in the plots.



**Throughput vs. Latency curves for 3 best configurations.**



**GPU Memory vs. Latency curves for 3 best configurations.**

The following table summarizes each configuration at the measurement that optimizes the desired metrics under the given constraints.

| **Model Config Name** | **Max Batch Size** | **Dynamic Batching** | **Total Instance Count** | **p99 Latency (ms)** | **Throughput (infer/sec)** | **Max GPU Memory Usage (MB)** | **Average GPU Utilization (%)** |
|---|---|---|---|---|---|---|---|
| resnet50_config_30 | 64 | Enabled | 4:GPU | 244.039 | 2291.78 | 3186 | 98.3 |
| resnet50_config_14 | 64 | Enabled | 2:GPU | 242.088 | 2288.31 | 1972 | 97.0 |
| resnet50_config_6 | 64 | Enabled | 1:GPU | 474.423 | 2230.09 | 1364 | 95.0 |
| resnet50_config_default | 128 | Disabled | 1:GPU | 217.318 | 1547.15 | 1401 | 73.0 |