Online Result Summary

Model: yolov8n

GPU(s): 1 x NVIDIA GeForce RTX 3060

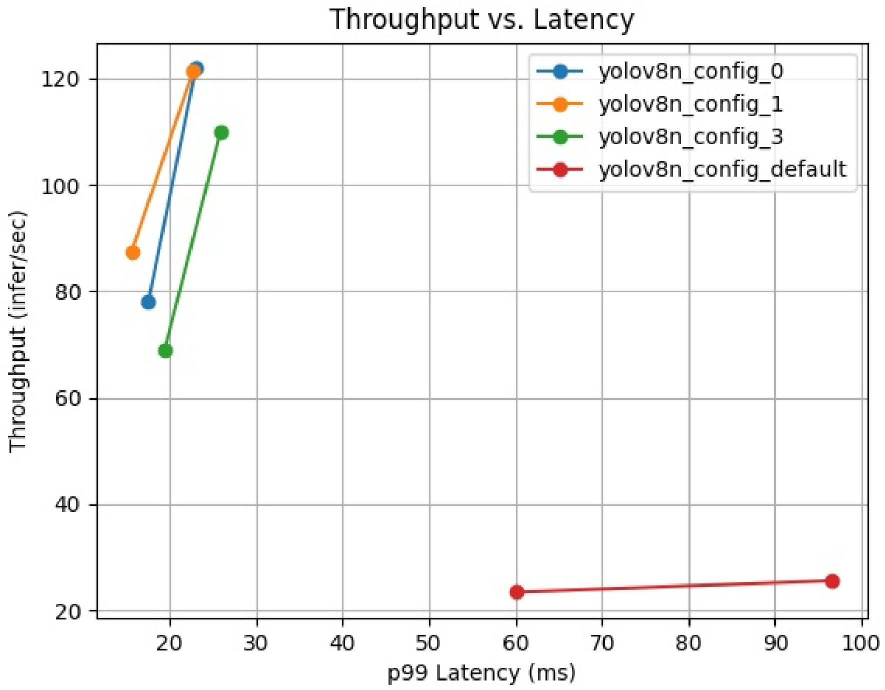
Total Available GPU Memory: 11.7 GB

Constraint targets: None

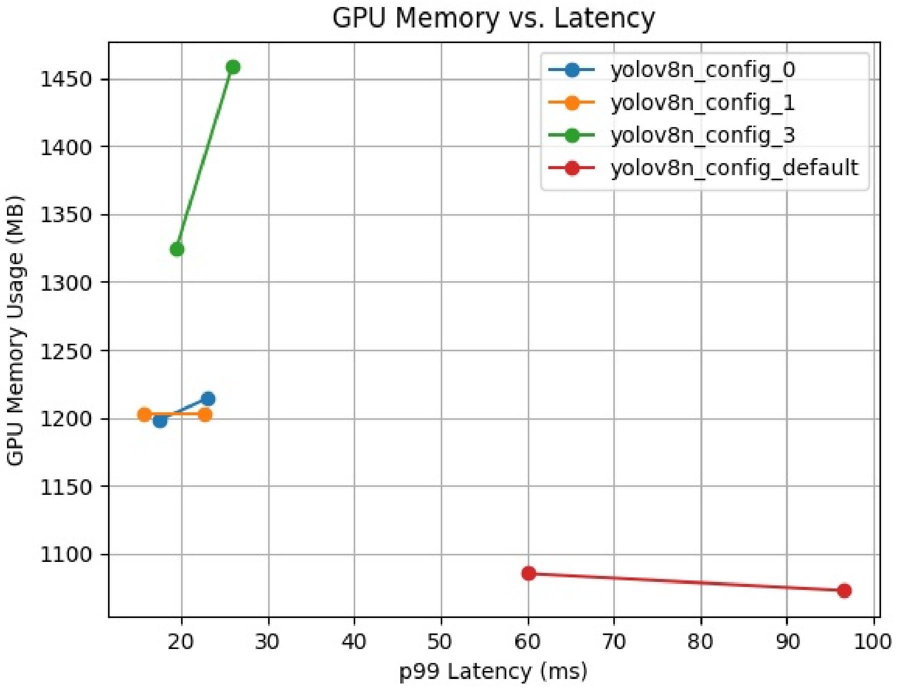
In 10 measurements across 5 configurations, **yolov8n_config_0** is **377%** better than the default configuration at maximizing throughput, under the given constraints, on GPU(s) 1 x NVIDIA GeForce RTX 3060.

- **yolov8n_config_0**: 1 GPU instance with a max batch size of 1 on platform onnxruntime

Curves corresponding to the 3 best model configuration(s) out of a total of 5 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 (%)
yolov8n_config_0	1	Enabled	1:GPU	23.035	121.852	1214	41.8
yolov8n_config_1	2	Enabled	1:GPU	22.706	121.254	1202	36.0
yolov8n_config_3	2	Enabled	2:GPU	25.891	109.879	1458	44.0
yolov8n_config_default	8	Enabled	1:CPU	96.648	25.5623	1072	27.8