Online Result Summary

Model: nnUnet_onnx

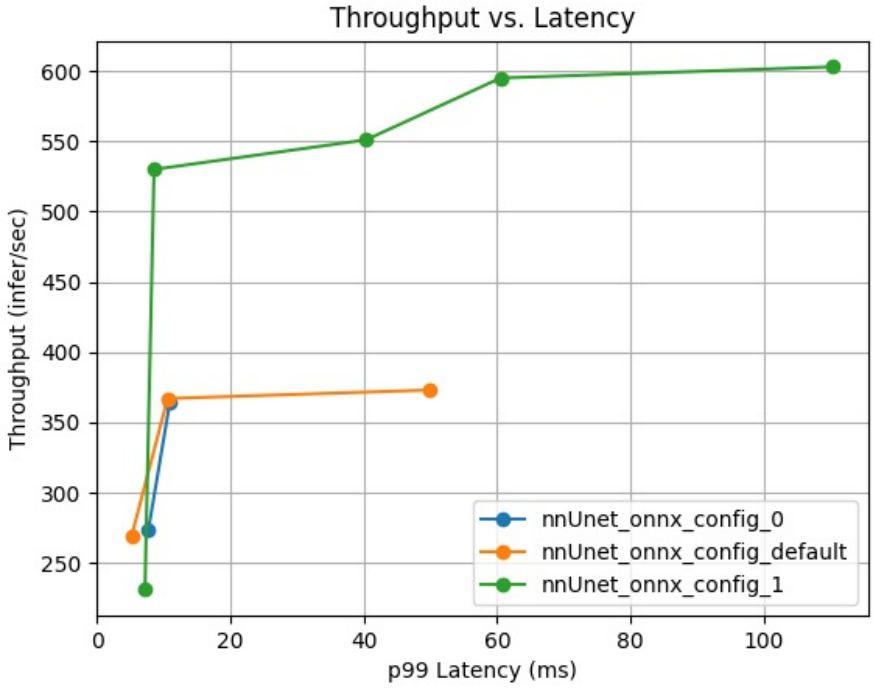
GPU(s): NVIDIA A100-PCIE-40GB

Total Available GPU Memory: 39.4 GB

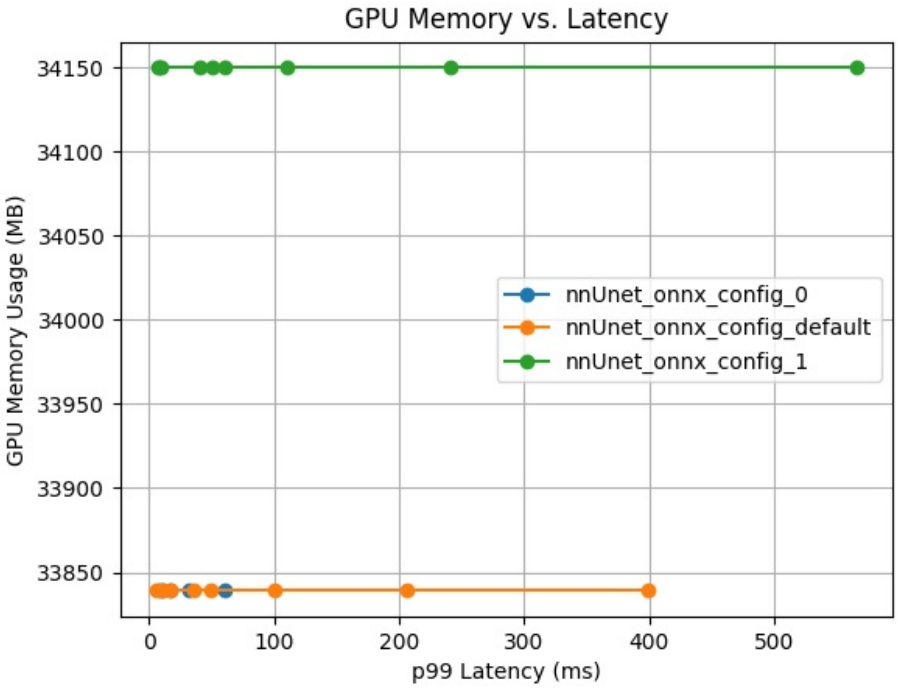
Constraint targets: None

In 50 measurement(s), config nnUnet_onnx_config_0 (1/GPU model instance(s) with max batch size of 1 and dynamic batching enabled) on platform onnxruntime_onnx delivers maximum throughput under the given constraints on GPU(s) NVIDIA A100-PCIE-40GB.

Curves corresponding to the 3 best model configuration(s) out of a total of 6 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.

	Max	Dynamic	Instance	p99	Throughput	Max CPU	Max GPU	Average GPU

Model Config Name	Batch Size	Batching	Count	Latency (ms)	(infer/sec)	Memory Usage (MB)	Memory Usage (MB)	Utilization (%)
nnUnet_onnx_config_0	1	Enabled	1/GPU	7.717	273.0	0	33839.0	21.6
nnUnet_onnx_config_default	1	Enabled	1/GPU	5.218	269.0	0	33839.0	23.8
nnUnet_onnx_config_1	1	Enabled	2/GPU	7.226	231.0	0	34150.0	20.9