Online Result Summary

Model: nnUnet

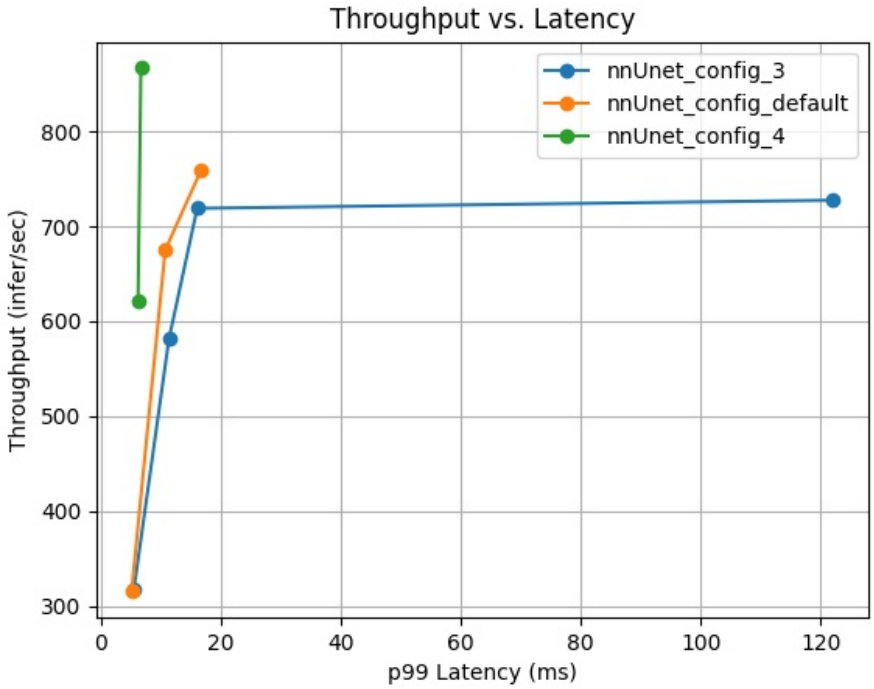
GPU(s): NVIDIA A100-PCIE-40GB

Total Available GPU Memory: 39.4 GB

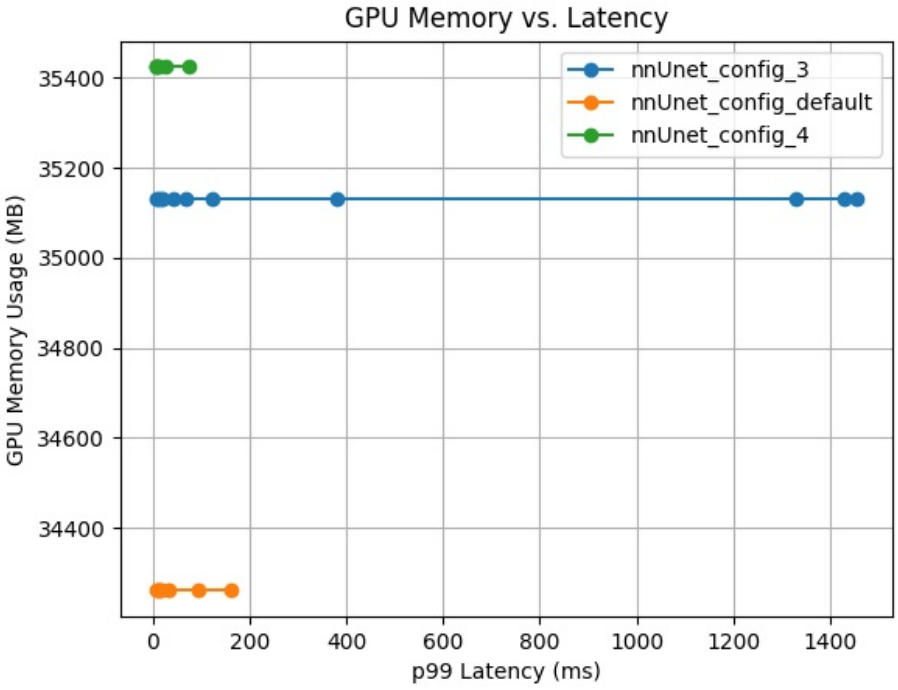
Constraint targets: None

In 52 measurement(s), config nnUnet_config_3 (4/GPU model instance(s) with max batch size of 1 and dynamic batching enabled) on platform tensorrt_plan 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			p99		Max CPU	Max GPU	Average

Model Config Name	Batch Size	Dynamic Batching	Instance Count	Latency (ms)	Throughput (infer/sec)	Memory Usage (MB)	Memory Usage (MB)	GPU Utilization (%)
nnUnet_config_3	1	Enabled	4/GPU	5.529	317.0	0	35131.0	18.2
nnUnet_config_default	1	Enabled	1/GPU	5.148	315.0	0	34261.0	21.0
nnUnet_config_4	1	Enabled	5/GPU	7.13	312.0	0	35424.0	20.0