Online Result Summary

Model: florence2_process

GPU(s):

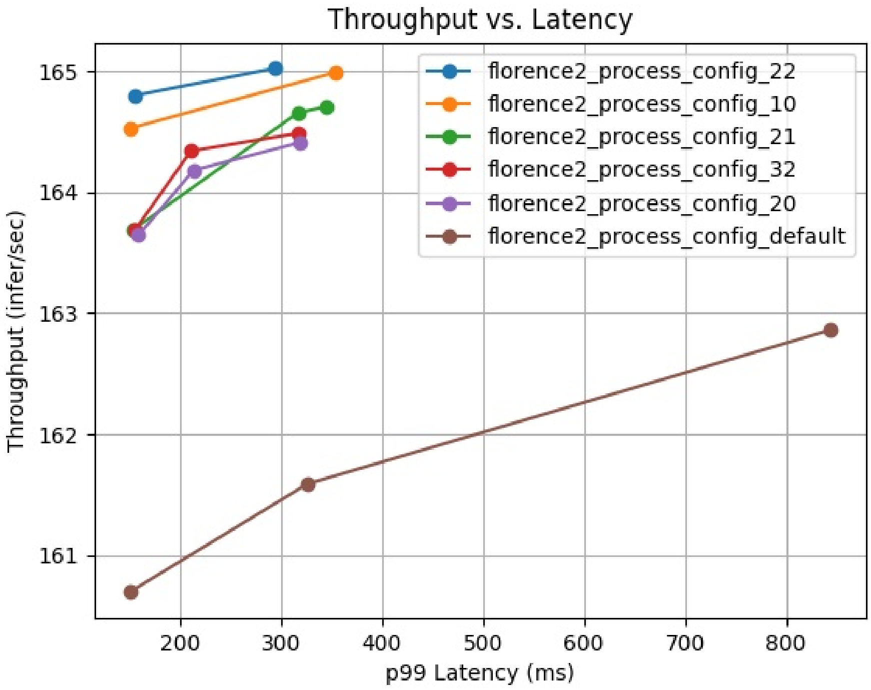
Total Available GPU Memory: 0 GB

Constraint targets: None

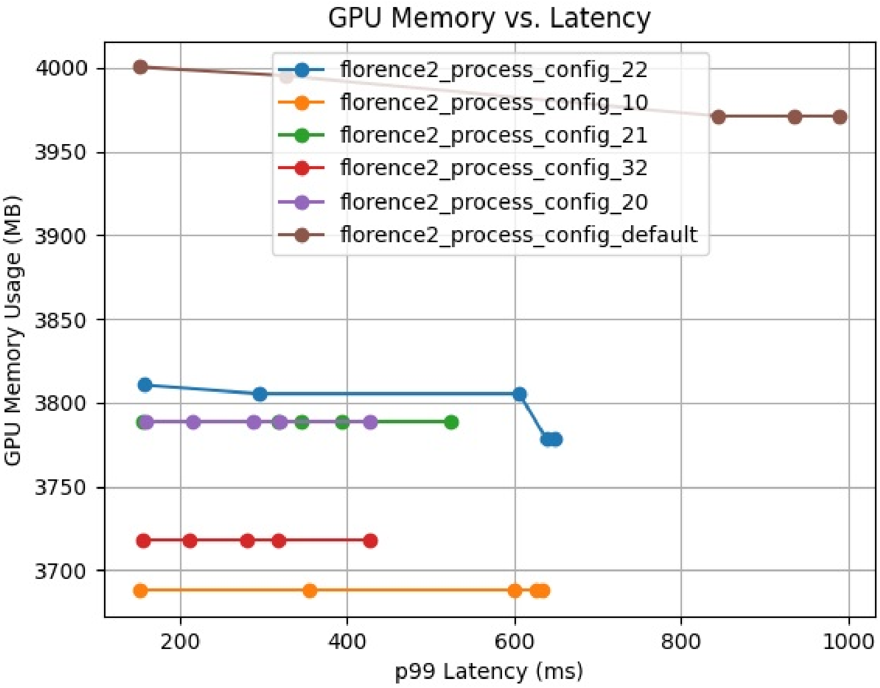
In 185 measurements across 37 configurations, **florence2_process_config_22** is **1%** better than the default configuration at maximizing throughput, under the given constraints, on GPU(s) .

- **florence2_process_config_22**: 4 CPU instances with a max batch size of 24 on platform python

Curves corresponding to the 5 best model configuration(s) out of a total of 37 are shown in the plots.



Throughput vs. Latency curves for 5 best configurations.



GPU Memory vs. Latency curves for 5 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 (%)
florence2_process_config_22	24	Enabled	4:CPU	294.141	165.025	3805	0.0
florence2_process_config_10	24	Enabled	4:CPU	354.15	164.992	3687	0.0
florence2_process_config_21	12	Enabled	4:CPU	345.599	164.708	3788	0.3
florence2_process_config_32	6	Enabled	4:CPU	317.005	164.488	3718	0.0
florence2_process_config_20	6	Enabled	4:CPU	318.761	164.411	3788	0.0
florence2_process_config_default	36	Enabled	4:CPU	844.385	162.862	3970	0.6