Online Result Summary

Model: wav2vec_py

GPU(s): 1 x NVIDIA GeForce GTX 1060 with Max-Q Design

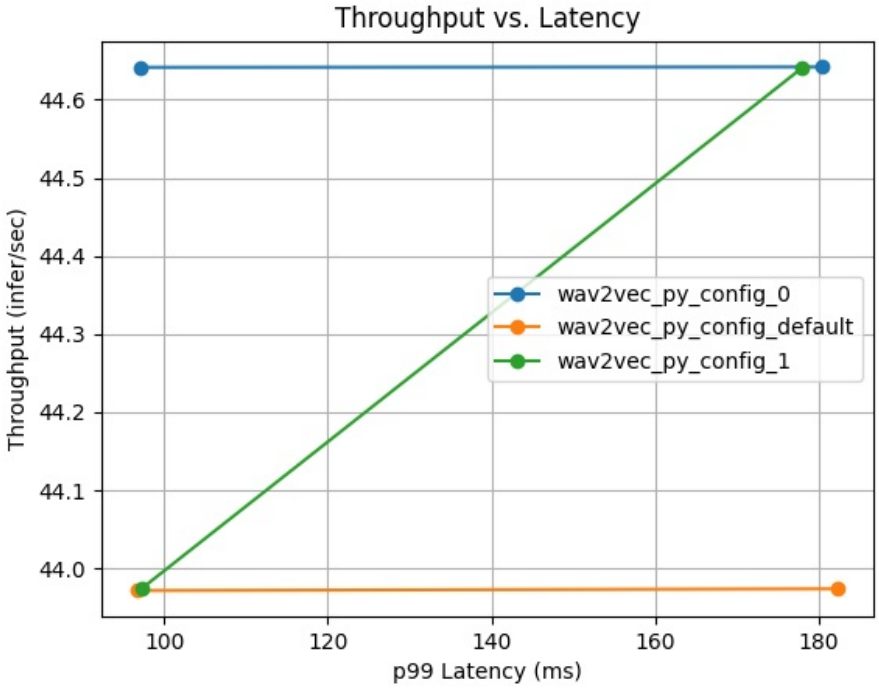
Total Available GPU Memory: 5.9 GB

Constraint targets: None

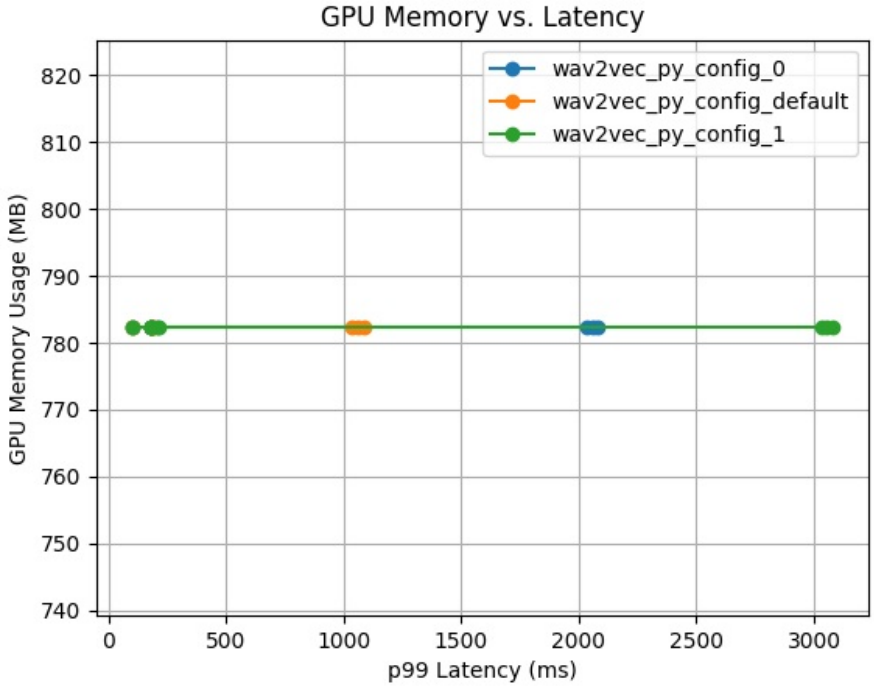
In 24 measurements across 3 configurations, **wav2vec_py_config_0** provides no gain over the default configuration, under the given constraints, on GPU(s) 1 x NVIDIA GeForce GTX 1060 with Max-Q Design.

- **wav2vec_py_config_0**: 1 GPU instance with a max batch size of 16 on platform python

Curves corresponding to the 3 best model configuration(s) out of a total of 3 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 (%)
wav2vec_py_config_0	16	Enabled	1:GPU	97.146	44.6415	782	81.9
wav2vec_py_config_default	16	Enabled	1:GPU	96.626	43.9716	782	79.0
wav2vec_py_config_1	16	Enabled	1:GPU	97.324	43.9748	782	70.7