Online Result Summary

Model: dexined_tf2

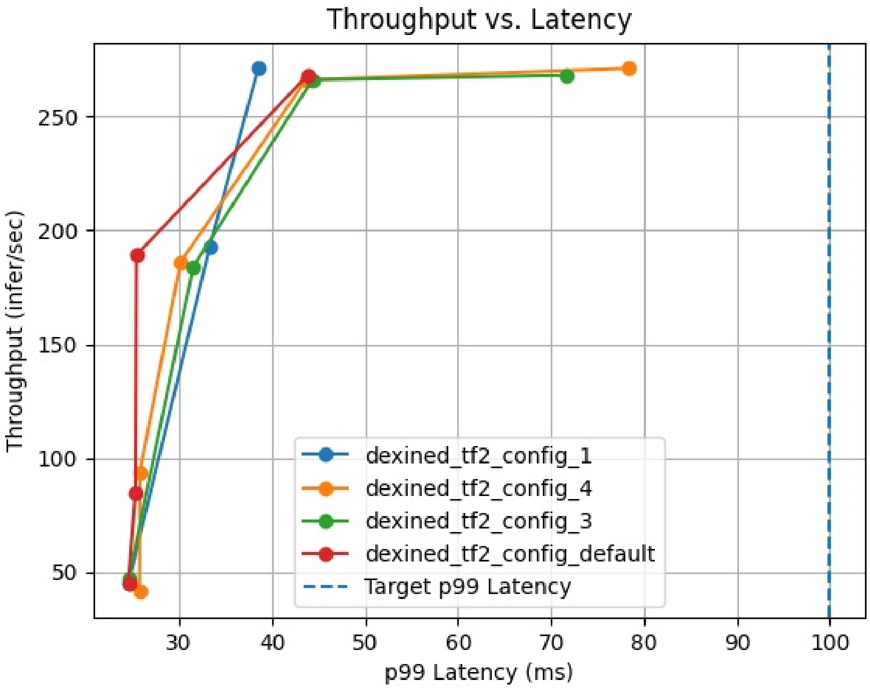
GPU(s): Tesla V100S-PCIE-32GB

Total Available GPU Memory: 31.7 GB

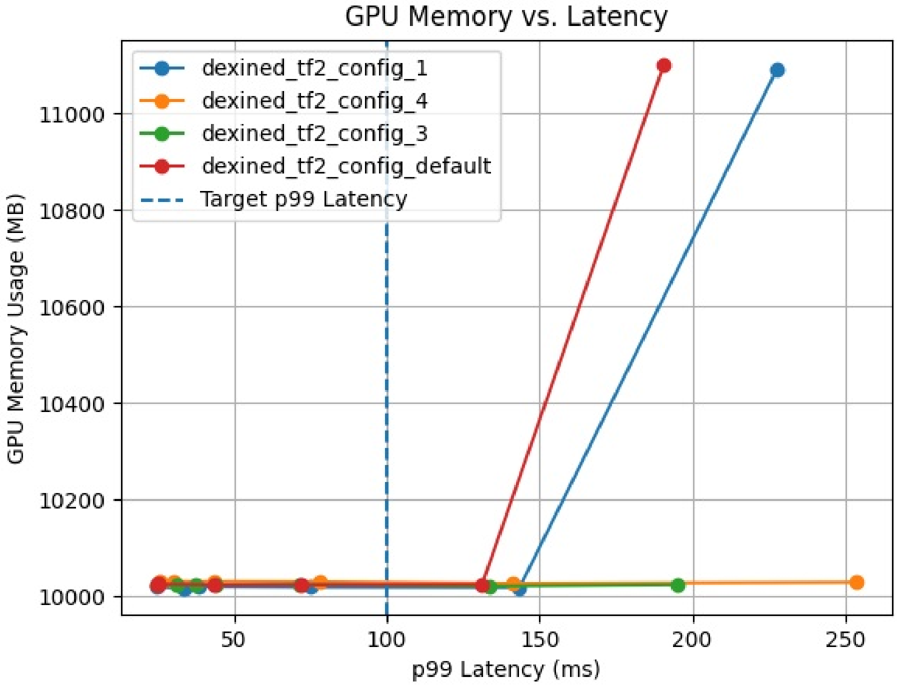
Constraint targets: Max p99 Latency : 100 ms

In 45 measurement(s), config dexined_tf2_config_1 (2/GPU model instance(s) with dynamic batching enabled) on platform tensorflow_savedmodel delivers maximum throughput under the given constraints on GPU(s) Tesla V100S-PCIE-32GB.

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.

			p99		Max CPU	Max GPU	Average

Model Config Name	**Dynamic Batching**	**Instance Count**	**Latency (ms)**	**Throughput (infer/sec)**	**Memory Usage (MB)**	**Memory Usage (MB)**	**GPU Utilization (%)**
dexined_tf2_config_1	Enabled	2/GPU	38.511	271.0	0	10020.0	72.0
dexined_tf2_config_4	Enabled	5/GPU	78.299	271.0	0	10030.0	37.3
dexined_tf2_config_3	Enabled	4/GPU	71.586	268.0	0	10025.0	40.8
dexined_tf2_config_default	Enabled	4/GPU	43.813	268.0	0	10023.5	71.1