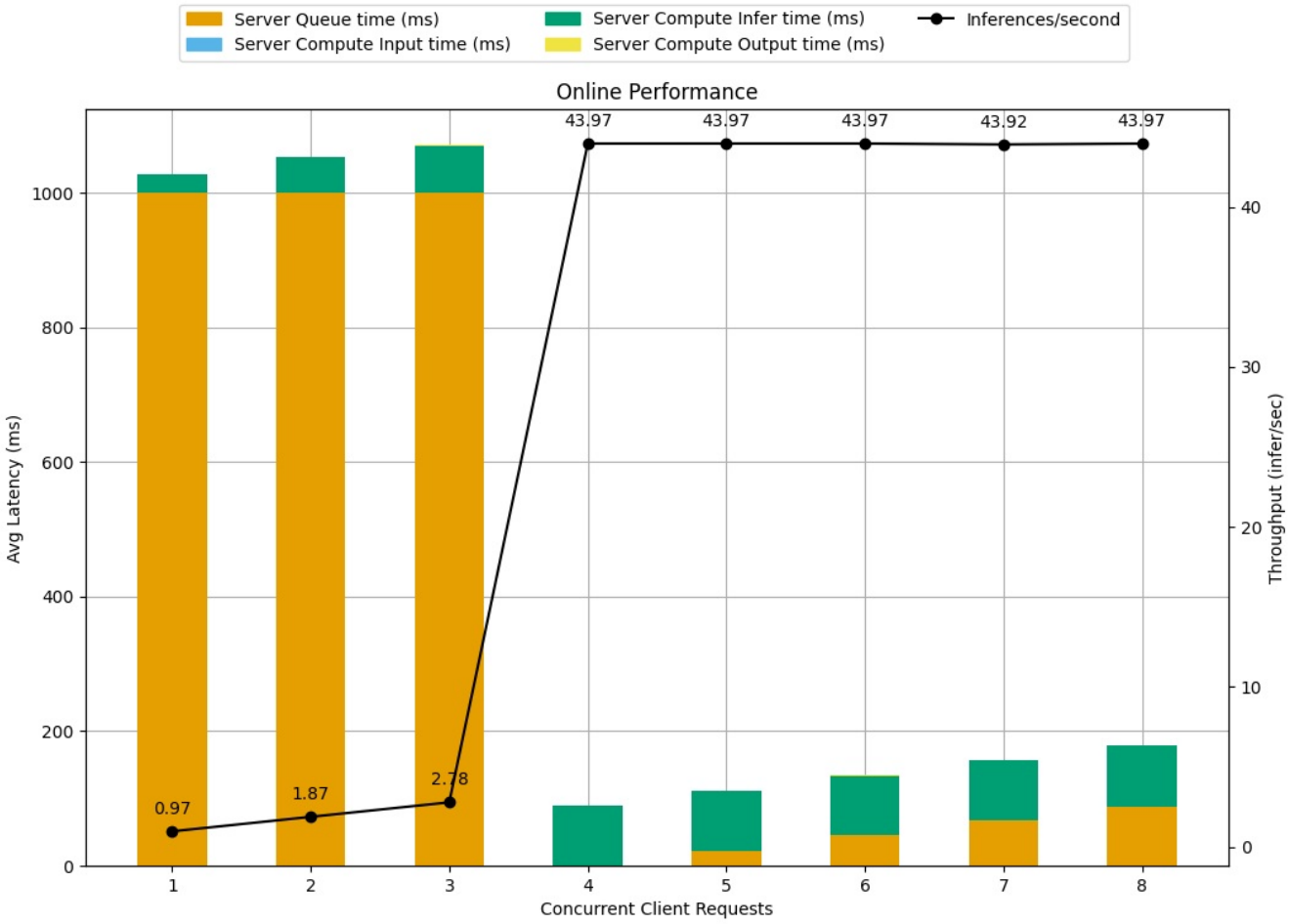
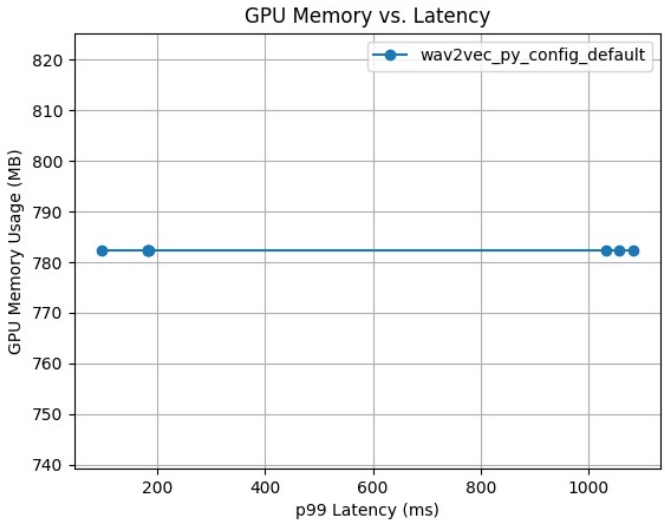
Detailed Report

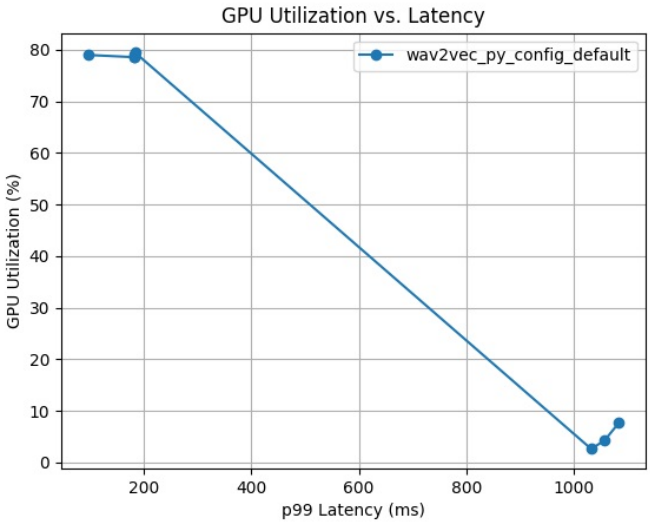
Model Config: wav2vec_py_config_default



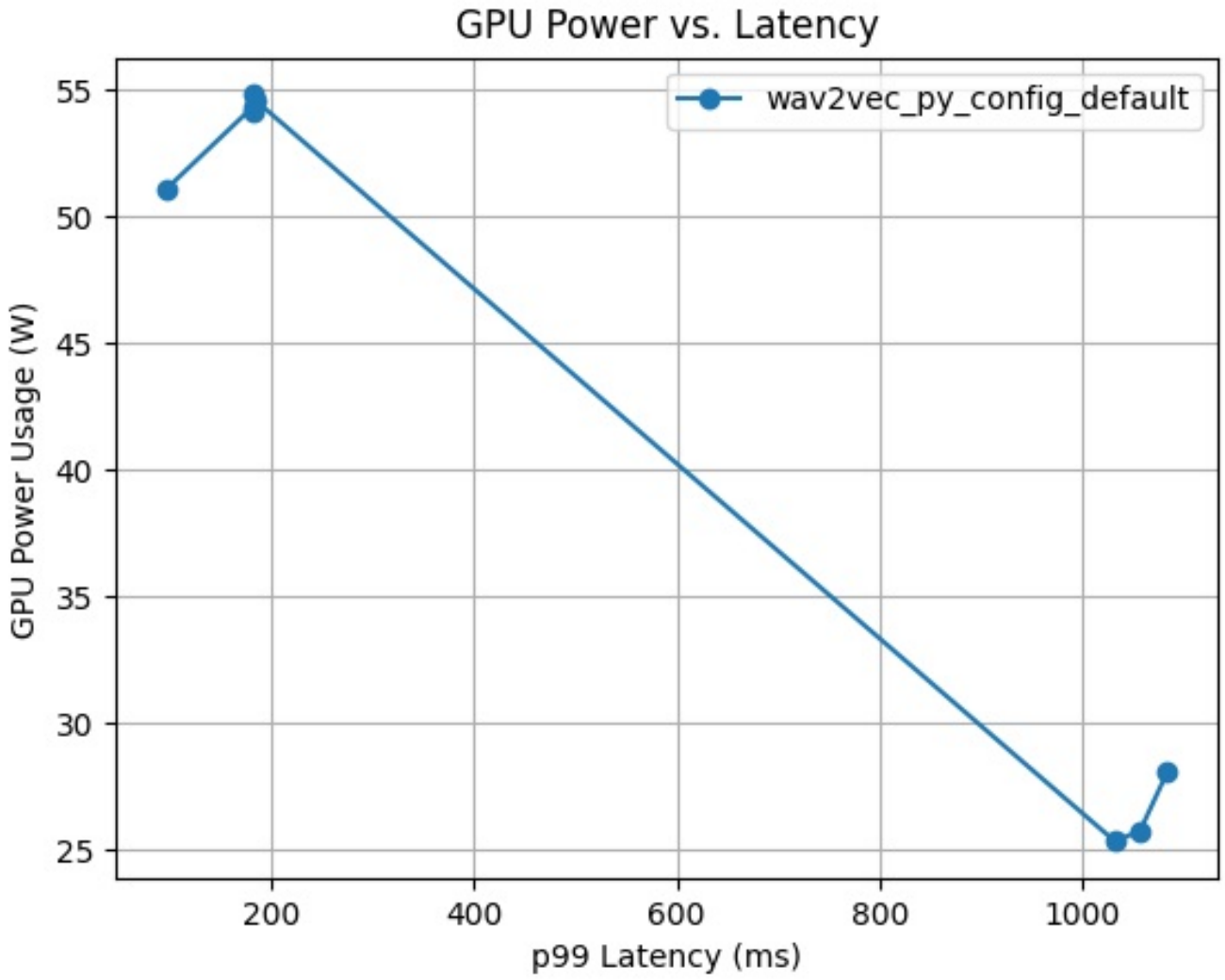
Latency Breakdown for Online Performance of wav2vec_py_config_default



GPU Memory vs. Latency curves for config wav2vec_py_config_default



GPU Utilization vs. Latency curves for config wav2vec_py_config_default



GPU Power vs. Latency curves for config wav2vec_py_config_default

Request Concurrency	p99 Latency (ms)	Client Response Wait (ms)	Server Queue (ms)	Server Compute Input (ms)	Server Compute Infer (ms)	Throughput (infer/sec)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
3	1083.569	1072.294	1000.102	0.147	70.898	2.78148	782.237696	7.7
2	1056.68	1054.884	1000.148	0.105	53.57	1.87479	782.237696	4.3
1	1031.87	1029.274	1000.241	0.073	28.068	0.967629	782.237696	2.6
8	184.695	179.638	88.295	0.179	89.897	43.9702	782.237696	79.4
7	183.418	157.714	67.183	0.134	89.273	43.9164	782.237696	79.1
6	183.356	135.017	44.773	0.121	88.988	43.973	782.237696	78.7
5	182.403	112.755	22.505	0.117	88.99	43.9738	782.237696	78.6
4	96.626	90.369	0.314	0.109	88.767	43.9716	782.237696	79.0

The model config **wav2vec_py_config_default** uses 1 GPU instance with a max batch size of 16 and has dynamic batching enabled. 8 measurement(s) were obtained for the model config on GPU(s) 1 x NVIDIA GeForce GTX 1060 with Max-Q Design with total memory 5.9 GB. This model uses the platform .

The first plot above shows the breakdown of the latencies in the latency throughput curve for this model config. Following that are the requested configurable plots showing the relationship between various metrics measured by the Model Analyzer. The above table contains detailed data for each of the measurements taken for this model config in decreasing order of latency.