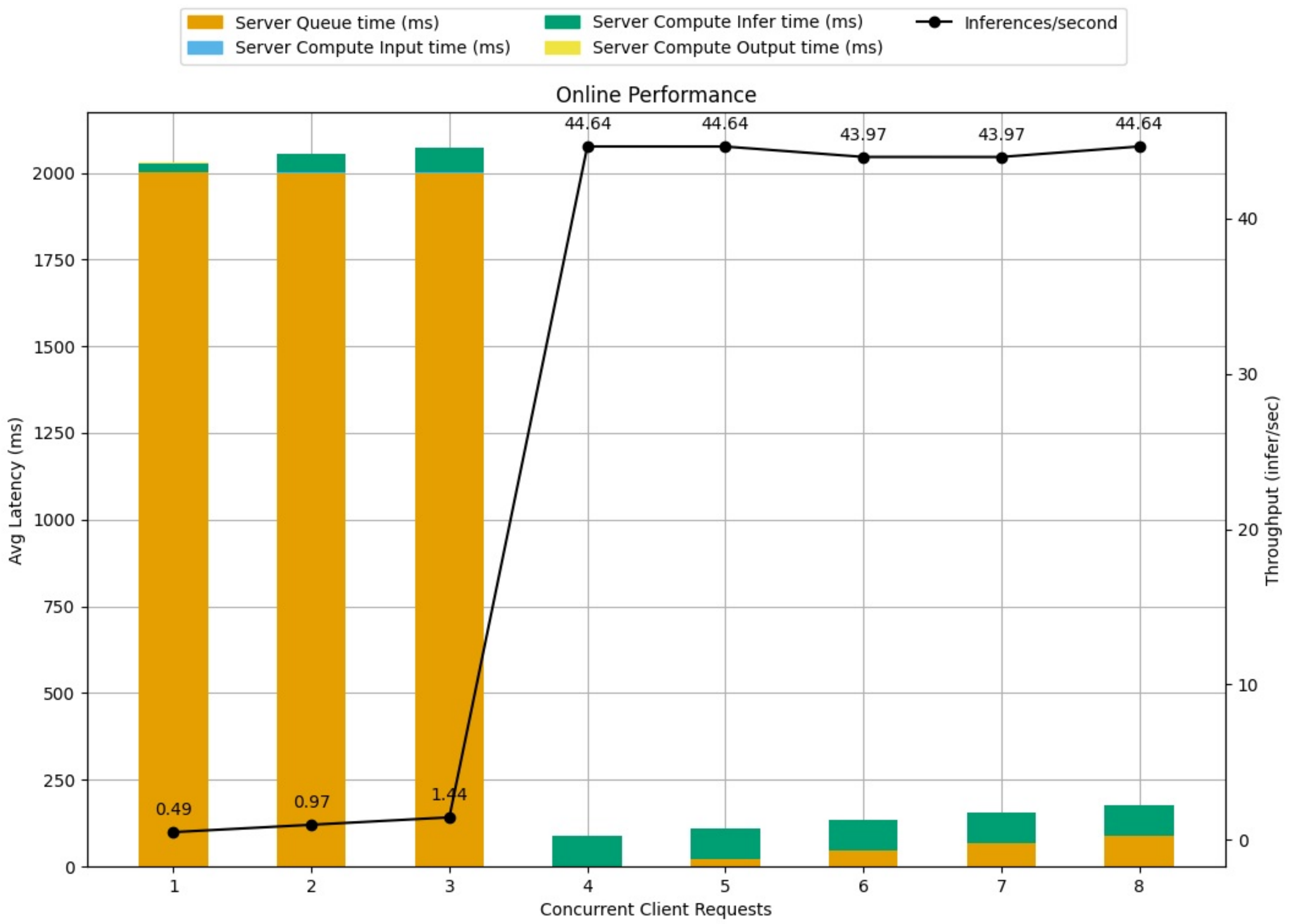
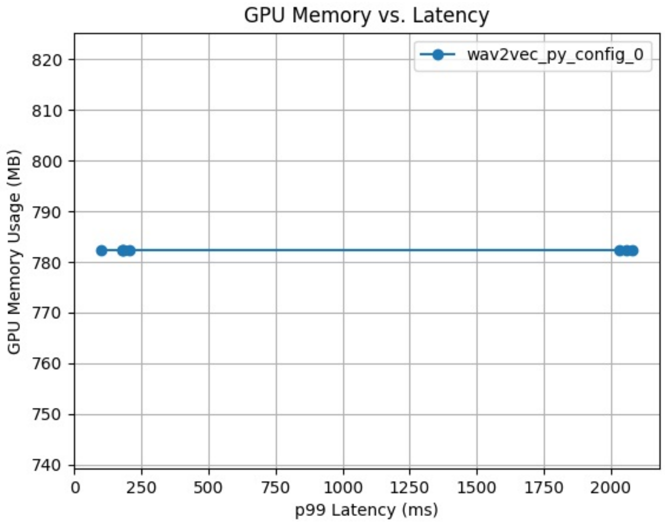
Detailed Report

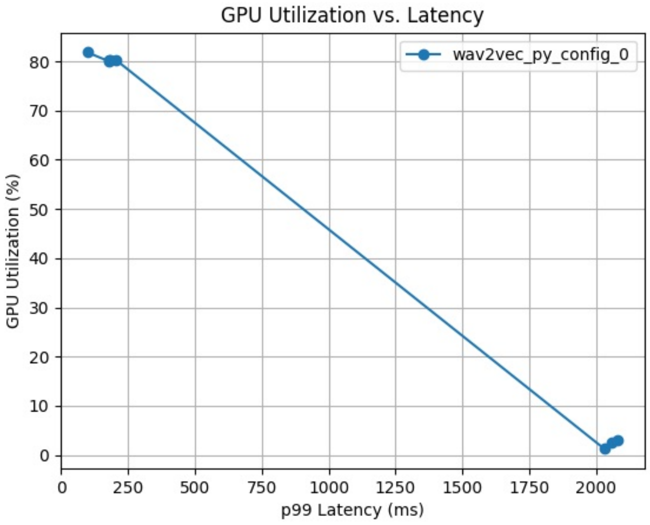
Model Config: wav2vec_py_config_0



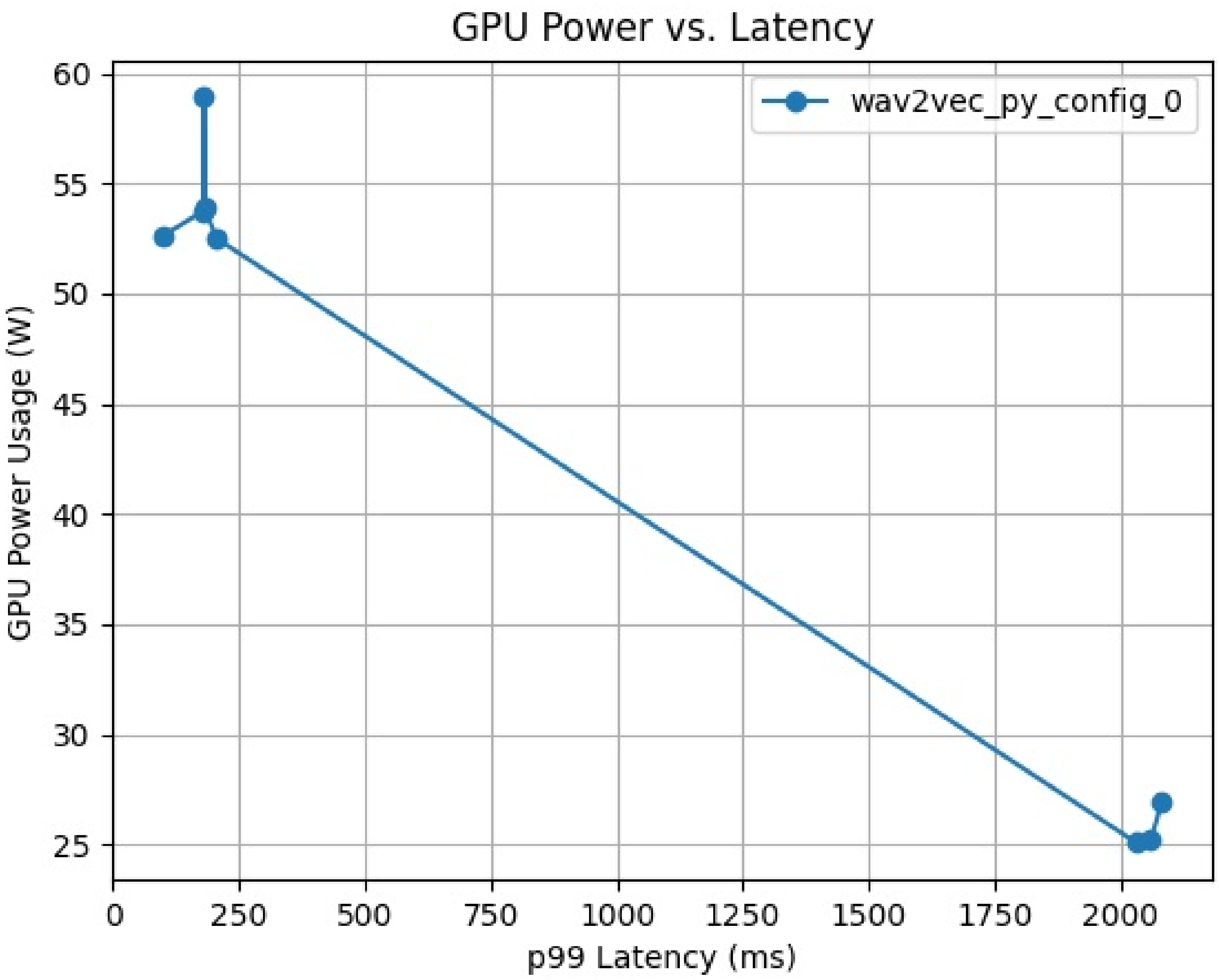
Latency Breakdown for Online Performance of wav2vec_py_config_0



GPU Memory vs. Latency curves for config wav2vec_py_config_0



GPU Utilization vs. Latency curves for config wav2vec_py_config_0



GPU Power vs. Latency curves for config wav2vec_py_config_0

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	2082.965	2072.052	2000.16	0.146	70.691	1.44316	782.237696	3.1
2	2061.003	2055.054	2000.176	0.105	53.737	0.967635	782.237696	2.5
1	2031.553	2029.231	2000.256	0.068	28.021	0.491751	782.237696	1.2
6	205.01	134.372	44.548	0.115	88.618	43.9683	782.237696	80.3
7	181.654	155.439	66.187	0.123	88.015	43.9653	782.237696	80.3
8	180.321	176.692	86.872	0.159	88.473	44.642	782.237696	80.3
5	179.766	110.916	22.138	0.111	87.56	44.6371	782.237696	80.0
4	97.146	88.964	0.213	0.102	87.539	44.6415	782.237696	81.9

The model config **wav2vec_py_config_0** 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.