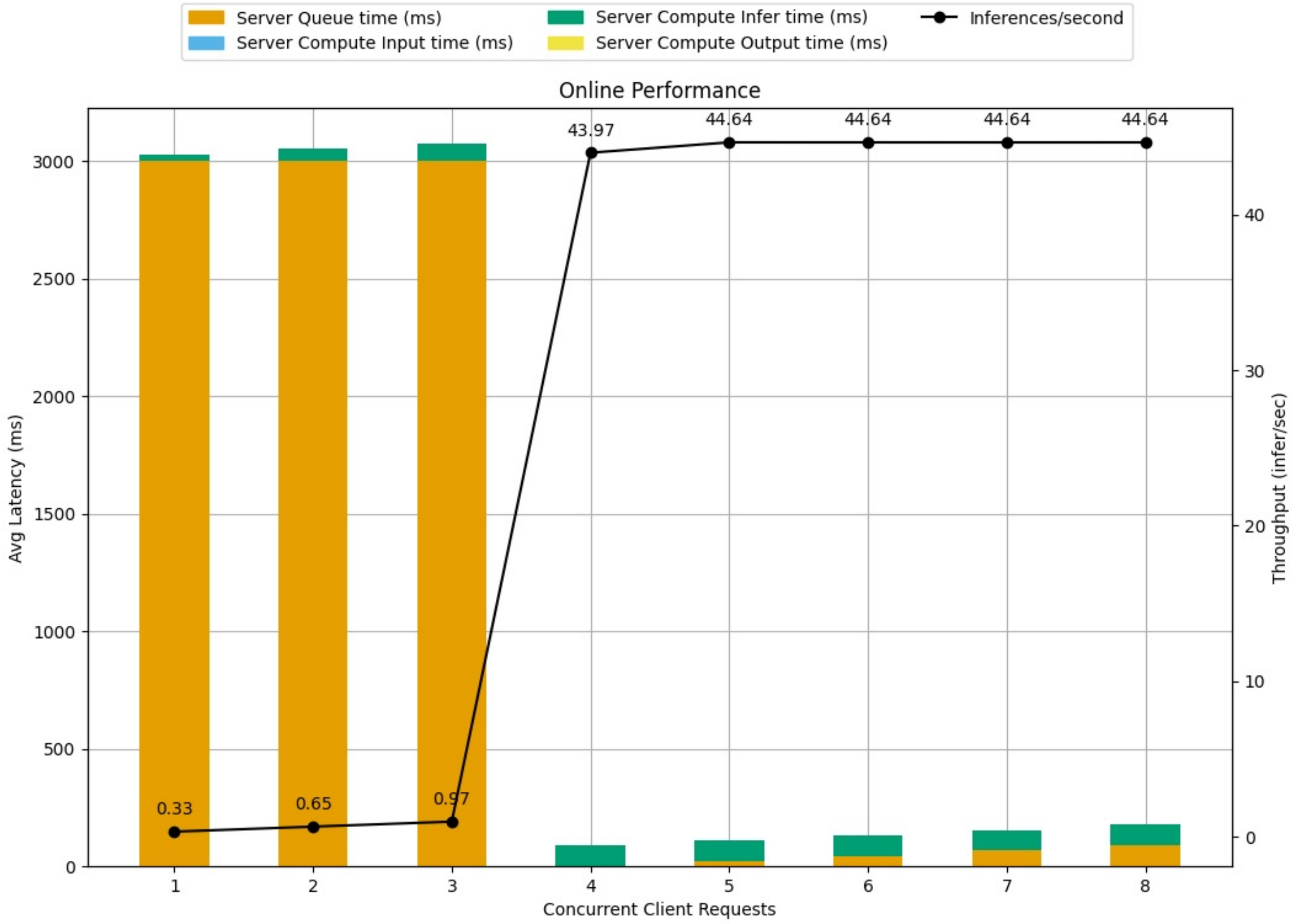
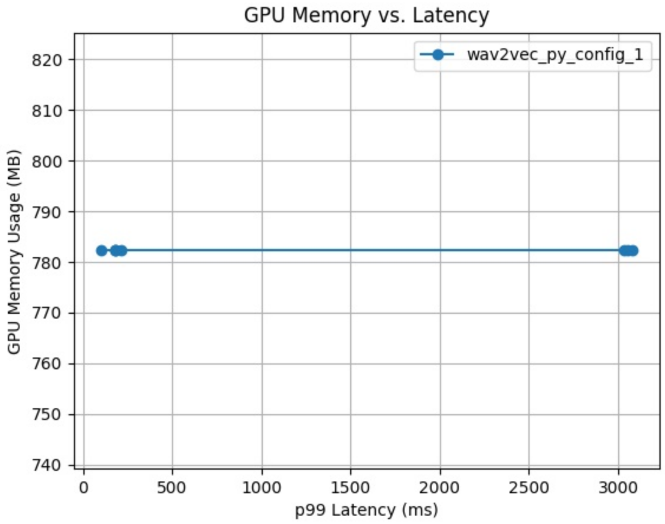
Detailed Report

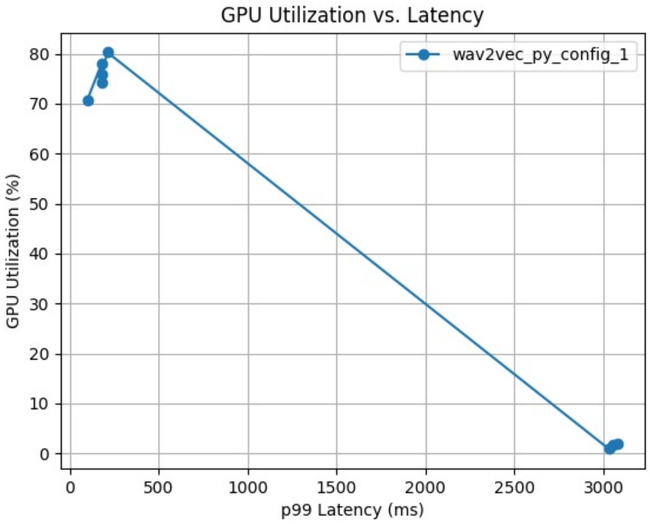
Model Config: wav2vec_py_config_1



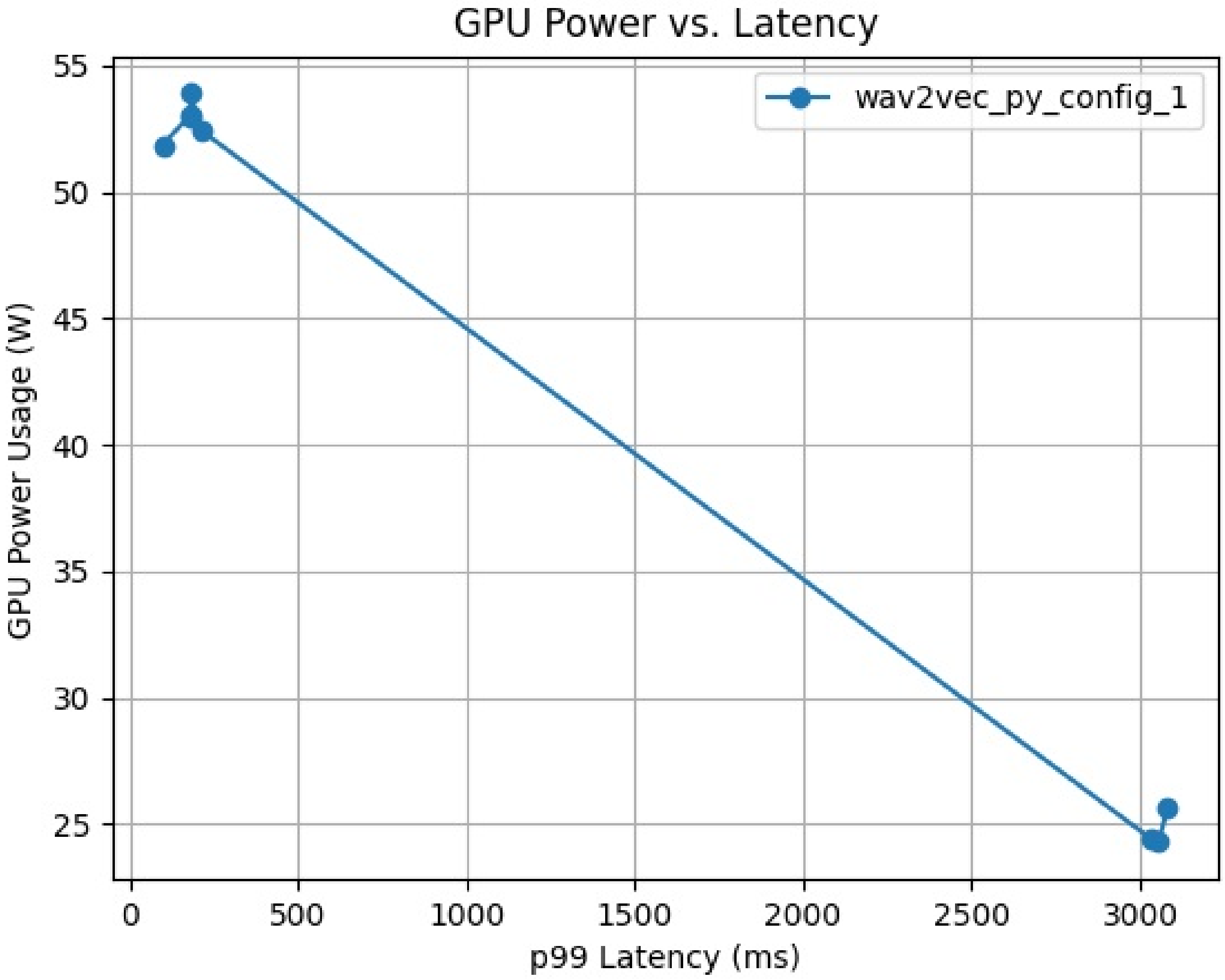
Latency Breakdown for Online Performance of wav2vec_py_config_1



GPU Memory vs. Latency curves for config wav2vec_py_config_1



GPU Utilization vs. Latency curves for config wav2vec_py_config_1



GPU Power vs. Latency curves for config wav2vec_py_config_1

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	3082.645	3072.684	3000.125	0.143	71.368	0.974383	782.237696	2.1
2	3055.578	3054.208	3000.231	0.114	52.928	0.652083	782.237696	1.8
1	3030.016	3028.721	3000.263	0.068	27.534	0.329622	782.237696	0.9
8	210.556	176.984	87.082	0.168	88.53	44.6384	782.237696	80.3
7	179.121	154.033	65.592	0.125	87.189	44.6375	782.237696	75.9
6	178.343	131.938	43.763	0.113	86.956	44.6396	782.237696	74.3
5	178.028	110.031	21.961	0.111	86.873	44.6412	782.237696	78.0
4	97.324	88.651	0.275	0.101	87.001	43.9748	782.237696	70.7

The model config **wav2vec_py_config_1** 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.