Design Space Exploration of Model Serving

Goal and Project Design: The goal of this project is to minimize the latency and energy consumption of model serving by optimizing the configuration space. We use average response time (ART) to measure latency, average cpu usage to measure energy consumption. We plan to test three models, image prediction (mnist), speech-to-text (deepspeech) and text translation. For the experiment infrastructure, we will have a load generator to send multiple requests, use TensorFlow serving to serve models and have docker to instantiate multiple serving boxes. The configuration space will focus on five variables; concurrency of requests, network reliability, CPU capacity, Memory capacity, number of servers.

Progress: So far, experimenting has been done on image prediction model mnist. Text Translation is ready for serving, while DeepSpeech model needs to be translated to TensorFlow model first. Below are result from the experiment on mnist model. We observed that the response times of first two inquiries are much higher than that of the rest, so we exclude them in following experiments. The average response time (ART) and CPU load is expected to increase with the increment of concurrency. But our current testing result show a drop in ART. We guess there may some cache operations happen in the serving model such that it takes less time to predict successive images. We test ART on different number of containers and find that 4 containers give the best performance. And then, we compare investigate the impact of CPU capacity and memory capacity on ART, find that ART drops when CPU/Memory capacity increases. These results were produced from running the experiment locally on a machine and we expect a general increase in response time when network reliability is decreased.

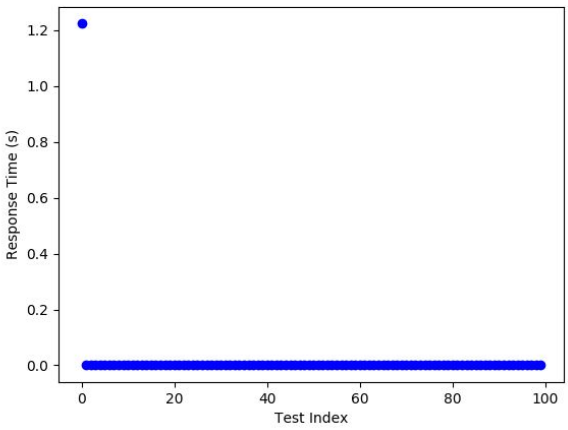

Fig 1. No outlier Excluded

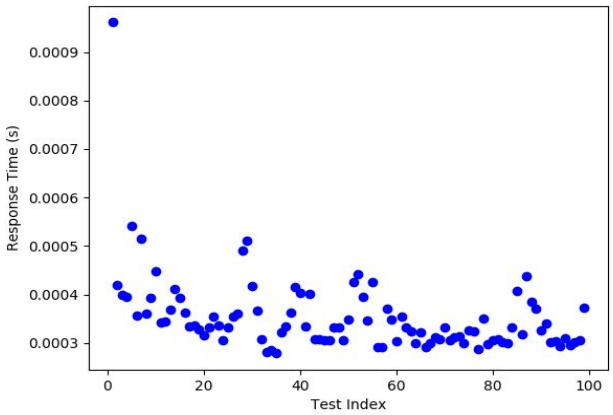

Fig 2 1st Outlier Excluded