Comparison of Go Web App and TensorFlow Serving API in Kubernetes

Tests were performed with the following values:

users	60000
spawn-rate	300
run-time	300s
user request frequency	1-2s

users - max concurrent users

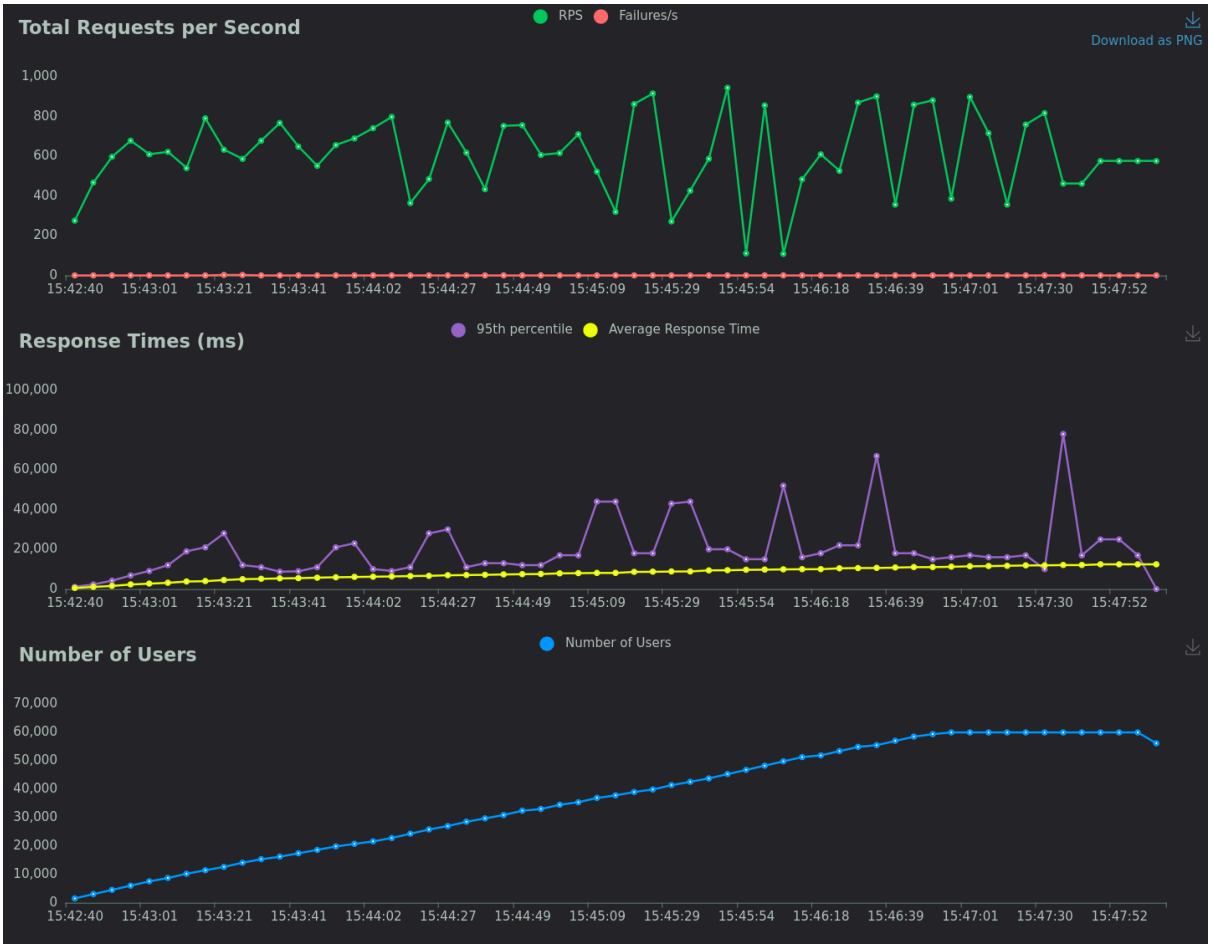
spawn-rate - number of new users each second

run-time - test duration

user request frequency - time between a user's requests

Webapp charts

Serving charts



General Statistics

Stats\Apps	Webapp	Serving
Requests Number	205477	194777
Median Response Time [ms]	7300	10000
Average Response Time [ms]	8802.02	12451.25
Last Throughput [requests per second]	714.19	575.7
Failed Requests	0	28

Comparative Analysis

Webapp manages more requests per second than serving and a rather shorter response time, which might reflect the lighter nature of its processing compared to the machine learning computations in serving. Both applications reached a maximum number of pods.