TorchServe Benchmarking Report

This document provides results from running benchmark for select models. Report focuses on latency and throughput results. The test setup other details are provided.

Tools and Configurations

Setup and Configurations

- Pull PyTorch ‘serve’ component – git clone https://github.com/pytorch/serve.git
- Setup - AWS EC2 node(s) - m4.xlarge (4 vCPU, 16GiB RAM)
- Install Python 3.7, Docker [latest] – Required by benchmark utility to spin new container for testing
- Install Jupyter and Pandas, Numpy, Metplotlib, Seaborn for analysis
- Create local image for torchserve. Use docker file Dockerfile.cpu under ‘dockers’ folder for serve component

Tools and Usage

- Git clone benchmark utility https://github.com/awslabs/deeplearning-benchmark.git
- Usage

  ```
  ./benchmark.sh –c 100 –n 1000 –w 4 --image ts_image\
      -u https://torchserve.s3.amazonaws.com/mar_files/resnet-18.mar
  ```

  Where, c – Concurrency, n - No. of Requests, image – Local docker image for testing

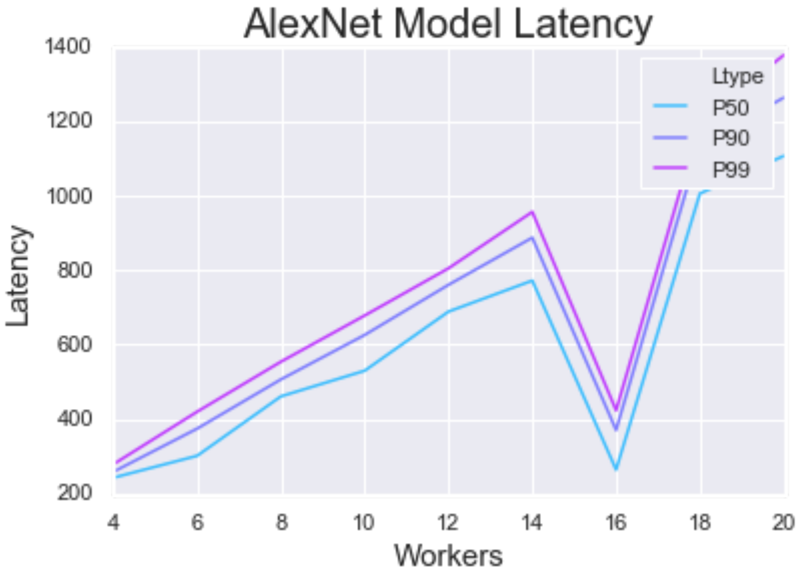
  For details refer README of the utility.
- Execution of above command results into following output –
  Where, P50, P90, P99 indicate model latency at 50th, 90th and 99th percentile of Requests respectively.

  ```
  <model details>
  <inference result>
  …
  TS version: torchserve == <your version>
  CPU/GPU: cpu
  Model: resnet-18
  Concurrency: 100
  Requests: 1000
  Model latency P50: 143.42
  Model latency P90: 146.26
  Model latency P99: 195.45
  TS throughput: 6.62
  TS latency P50: 14900
  TS latency P90: 15456
  TS latency P99: 15705
  TS latency mean: 15108.998
  TS error rate: 0.000000%
  ```

Latency Results

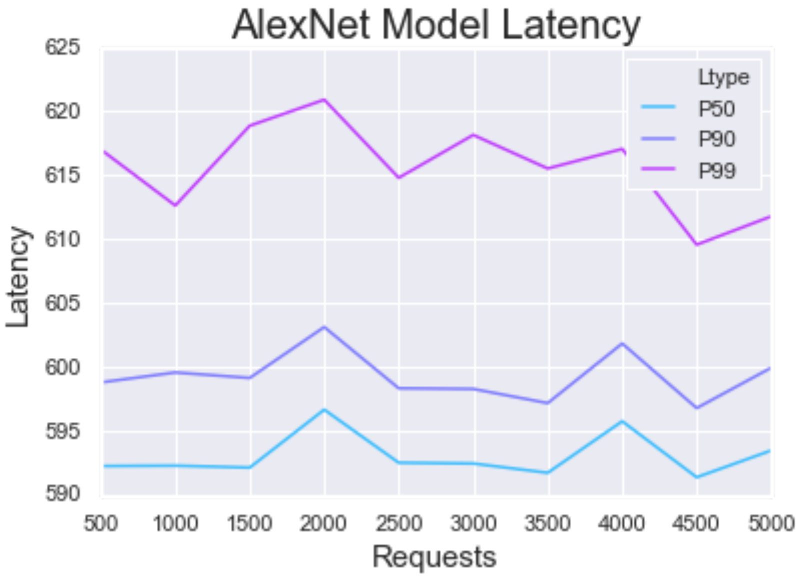
Alexnet Model

- Config – n=1000, c=10 and vary workers with [-w] from 4 to 20



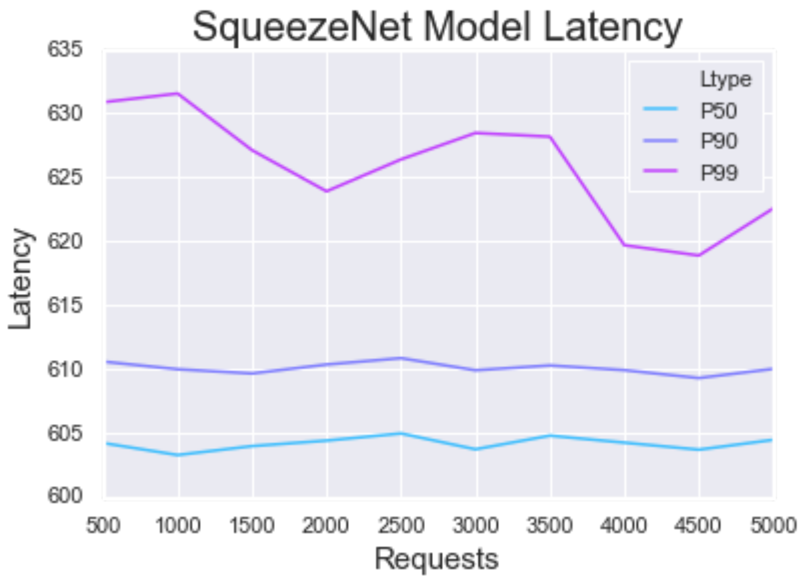
AlexNet Model

- Config. - w=4, c=10 and vary requests from 500 to 5000



SqueezeNet_v1.1 Model

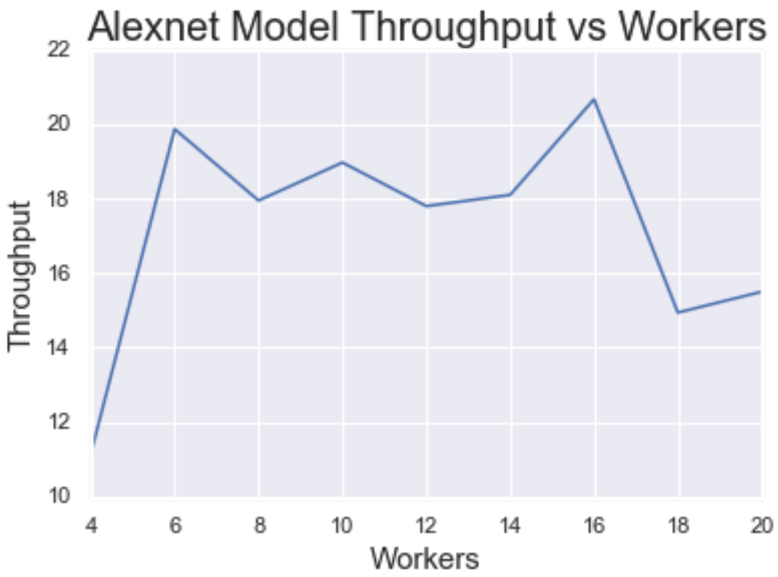
- Config. - w=4, c=10 and vary requests from 500 to 5000



Throughput Results

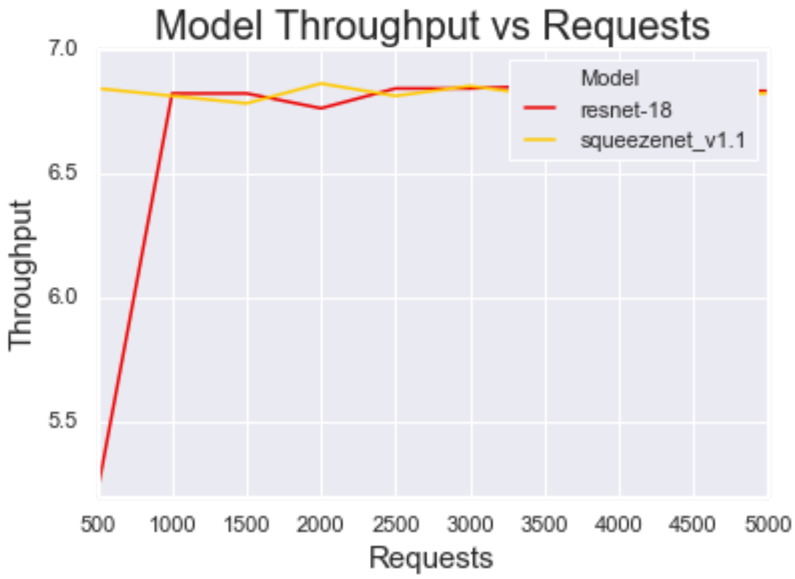
AlexNet Model

- Config. - n=1000, c=10 and vary workers with [-w] from 4 to 20



ResNet-18 and SqueezeNet_v1.1 Model

- Config. - w=4, c=10 and vary requests from 500 to 5000


Model Throughput vs Requests
Throughput
Requests
Model
resnet-18
squeezenet_v1.1