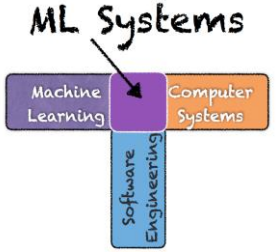
P3 – Design Space Exploration of Model Serving



Mid-Term Progress Report

Jianhai Su, Harrison Howell and Bharat Joshi

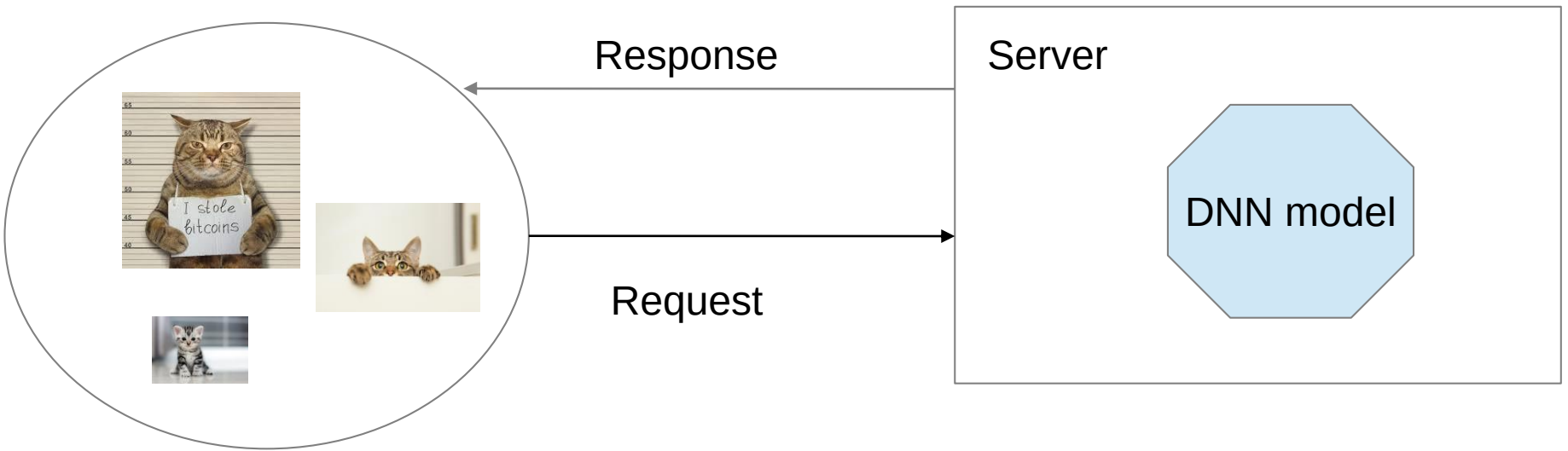
UNIVERSITY OF SOUTH CAROLINA

Outline

- Problem to Investigate
- Configuration Space
- Experimental Infrastructure
- DNN Models
- Investigation Plan
- Current Experimental Progress
- Following Steps

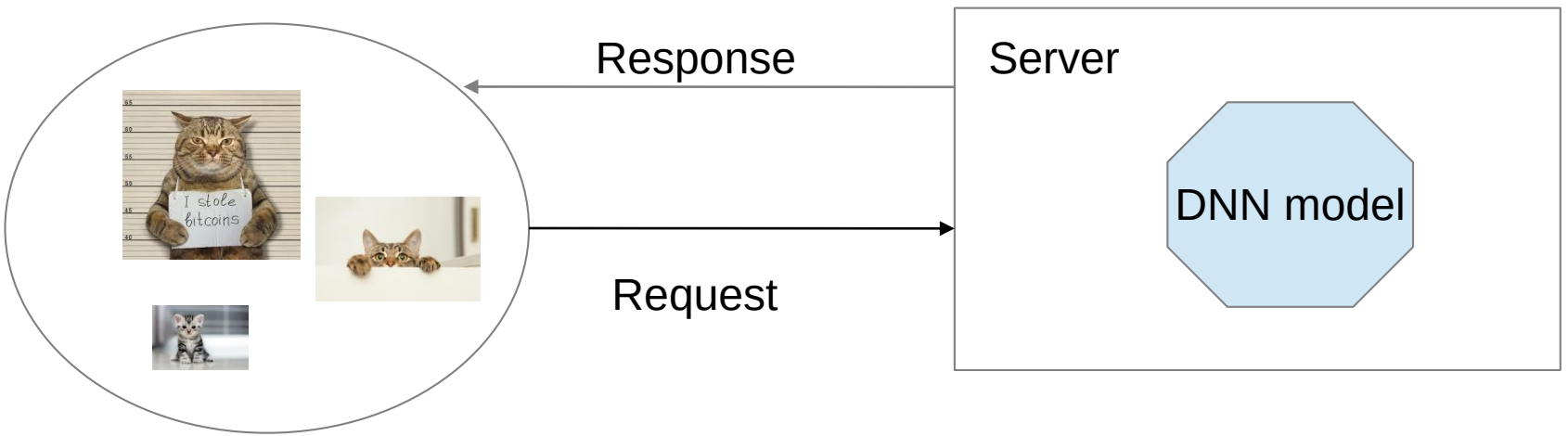
Problem to Investigate

- Latency of a request
- Energy Consumption of a serving model
- Configuration Space



Problem to Investigate

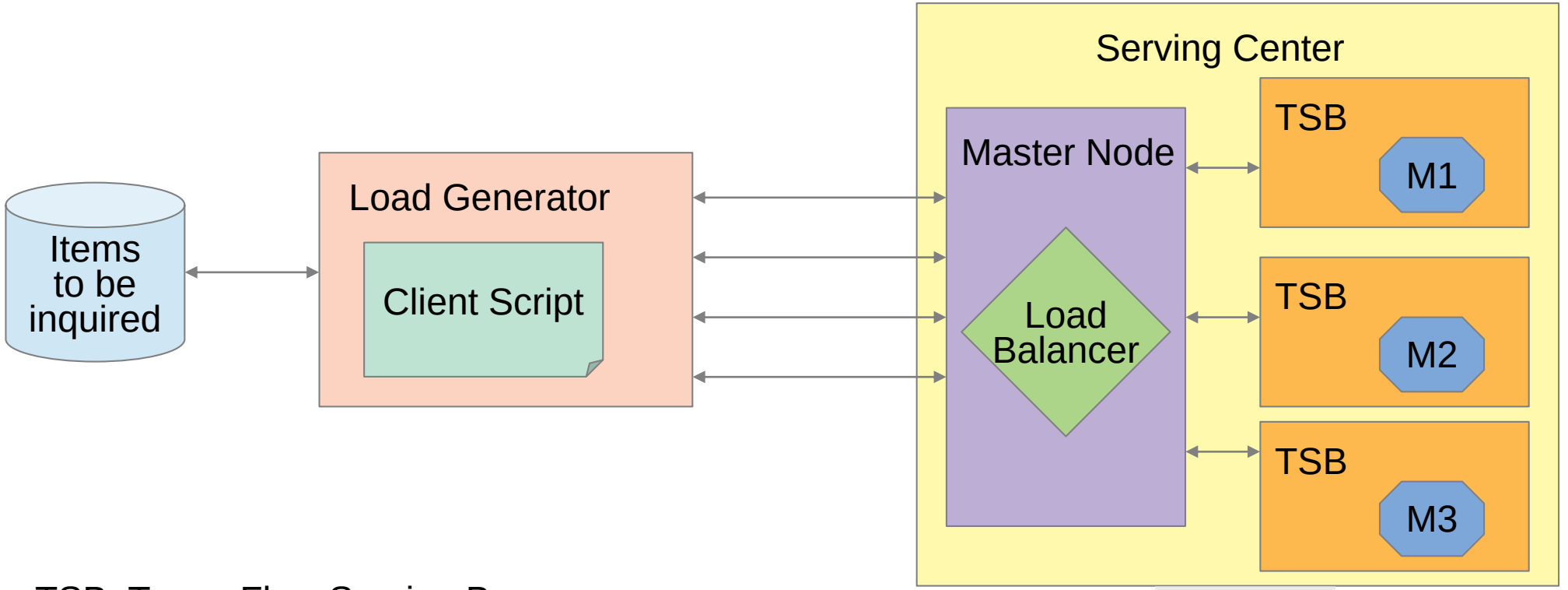
- Latency of a request: **average response time** (ART)
- Energy Consumption: **average CPU usage** (ACU)
  - Event Trigger vs Window Cut
  - psutil python lib: running **processes** and **system utilization**



Configuration Space

- Client Side
  - # of threads
  - Size of a file to be inquired
- Server Side
  - # of servers
  - CPU capacity
  - Memory capacity
- Network reliability:?

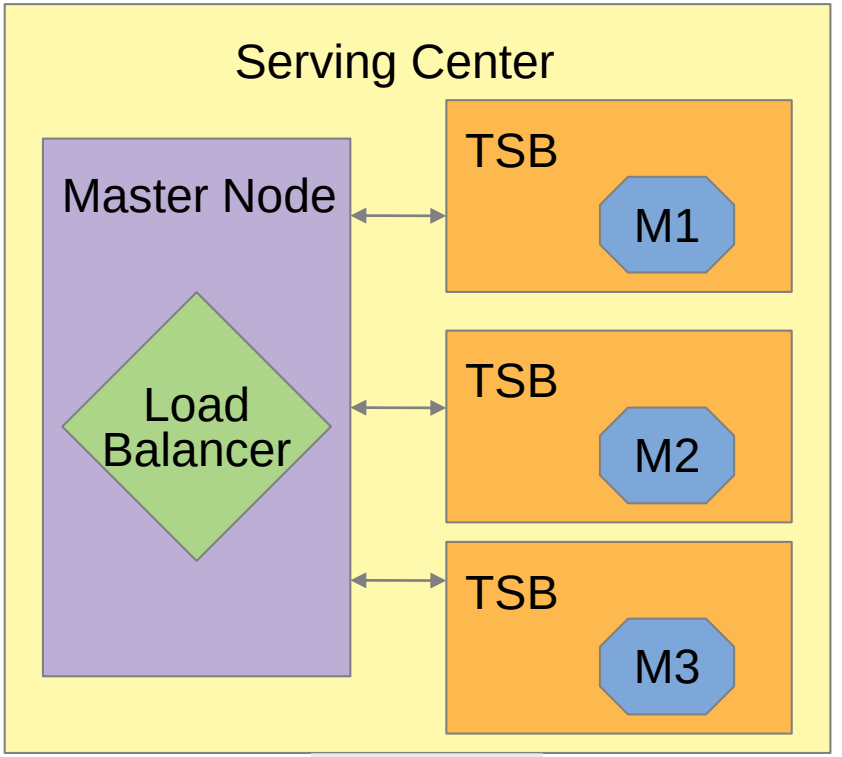
Experimental Infrastructure



TSB: TensorFlow Serving Box

Experimental Infrastructure

```
File  Edit  View  Search  Terminal  Help
1 docker-compose.yml
version: "3"
services:
  web:
    # replace username/repo:tag with your name and image details
    image: oceank/image_prediction_serving:mnist
    deploy:
      replicas: 4
      resources:
        limits:
          cpus: "0.2"
          memory: 1000M
      restart_policy:
        condition: on-failure
    ports:
      - "8500:8500"
    networks:
      - webnet
networks:
  webnet:
~
~
~
~
~
~
~
~
~
~
~
~
```



DNN Models

- Image: Inception V3 (CNN), Simple Softwmax Regression
- Text Translation: Neural Machine Translation (RNN)
- Speech-To-Text: DeepSpeech V0.1.0 (RNN)

Investigation Plan

- Impact of single configurable parameter
  - # of threads, size of file to be inquired
  - # of servers, CPU capacity, Memory capacity
- Correlation of 5 parameters on the impact of ART and ACU

Current Experimental Progress

- Servable Model: simple softmax regression model
  - Trained on MNIST dataset

- MNIST dataset
  - handwritten digits
  - Black-white
  - size 28X28
  - Testing samples: 10k

☆Another servable model: Neural Machine Translation Model

Current Experimental Progress

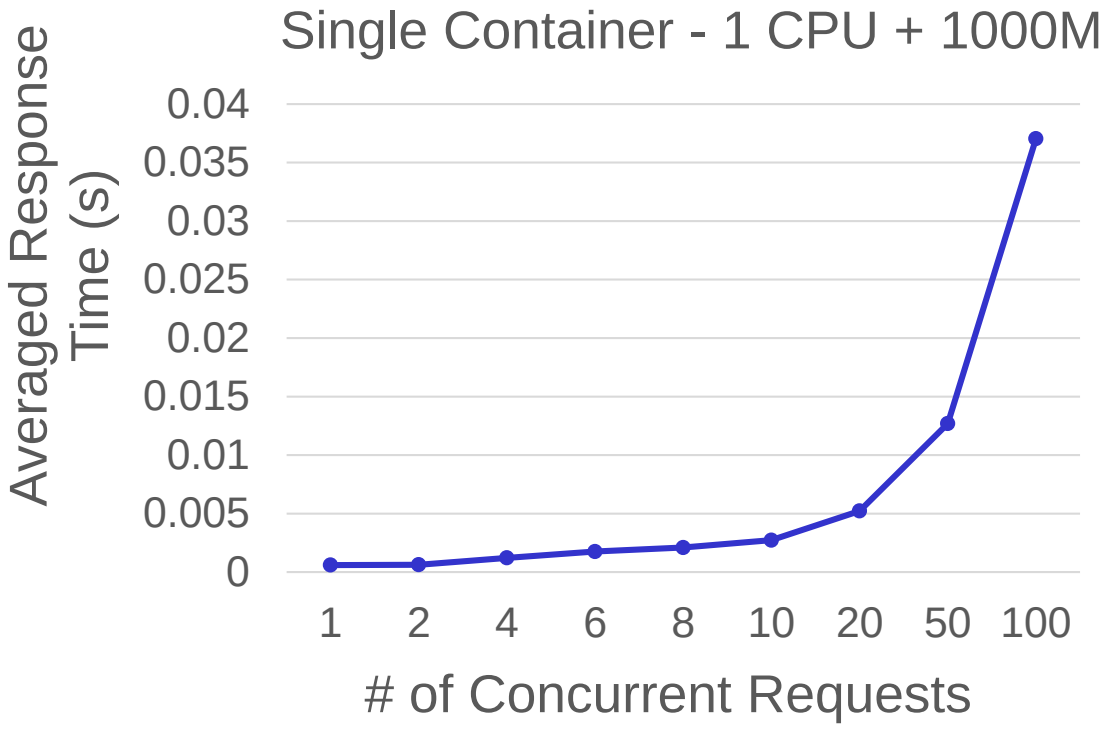
Impact of # of concurrent requests

ART **increases** when # of concurrent requests increases

# of Tests: 1000

# of Containers: 1

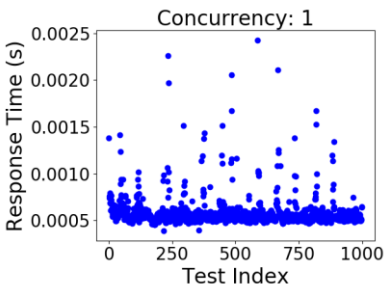
For each container:
CPU: 1.0 core
Memory: 1000M

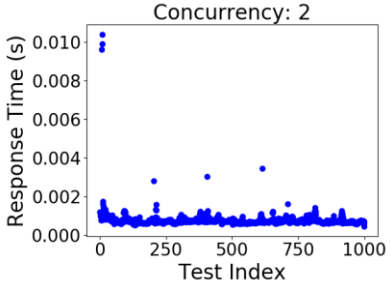


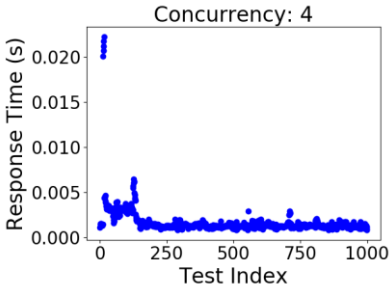
Current Experimental Progress

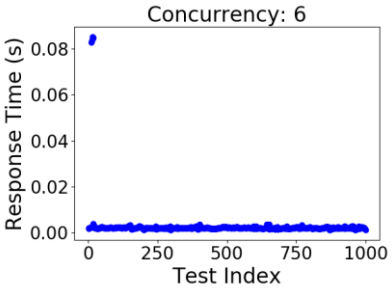
Model: warm up & Cool down

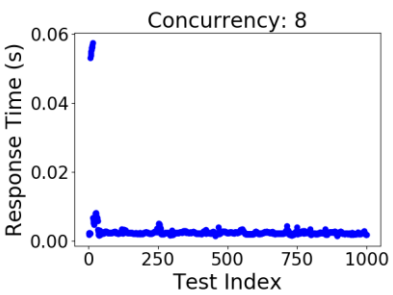
When the # of concurrent requests increases, windows of model warm-up and model cool-down appear.

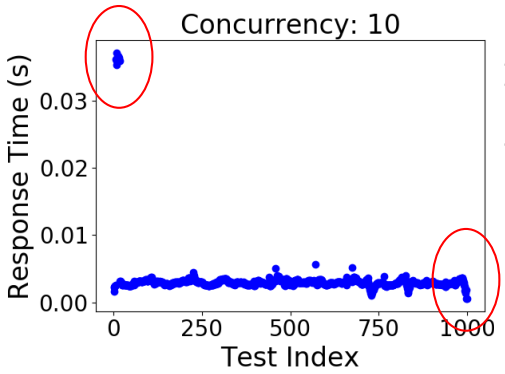


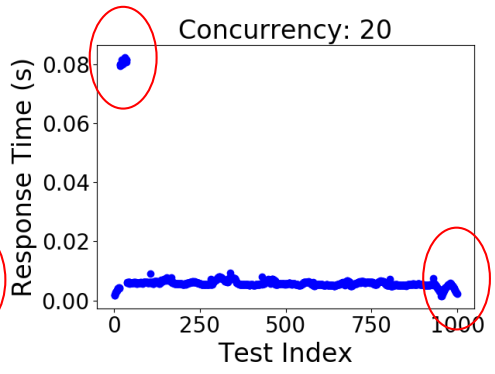


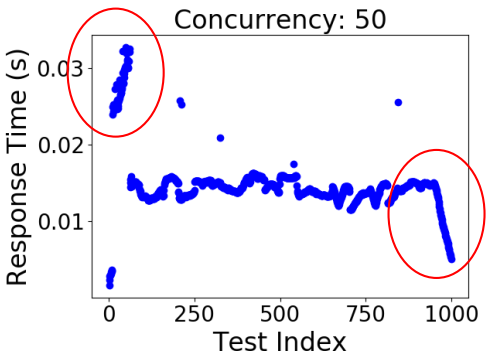


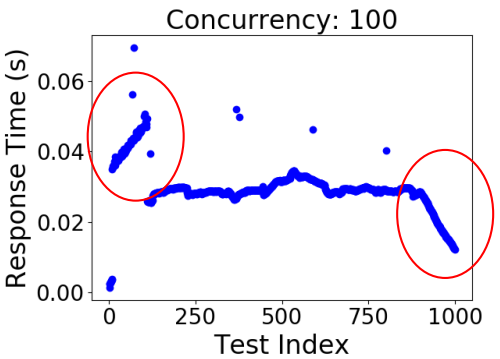












Current Experimental Progress

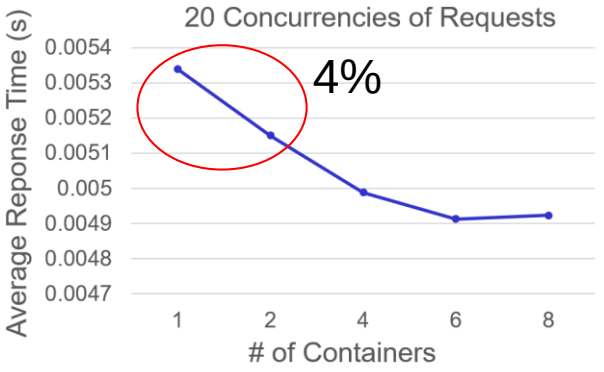
Impact of # of Containers

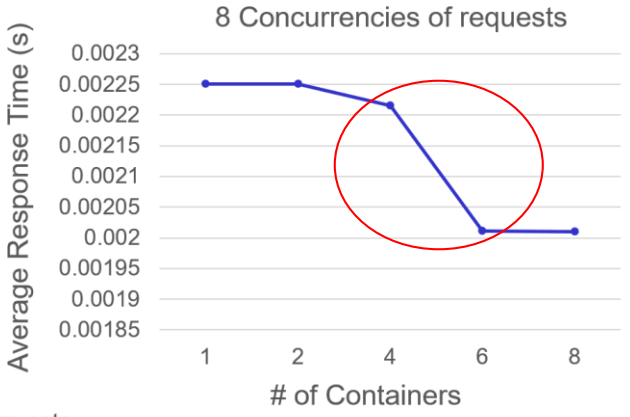
ART drops in general when # of servers increases

# of Tests: 1000

For each container:
CPU: 1.0 core
Memory: 1000M

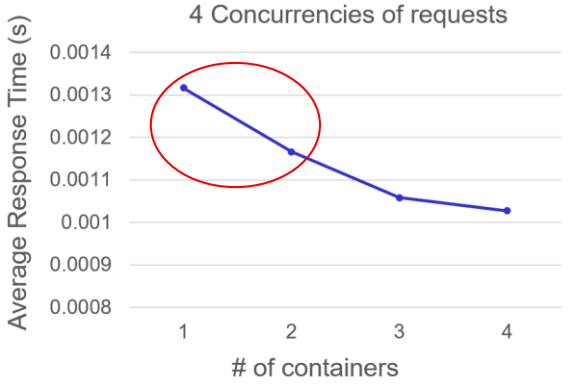
Desktop:
8 cores
2 threads per core





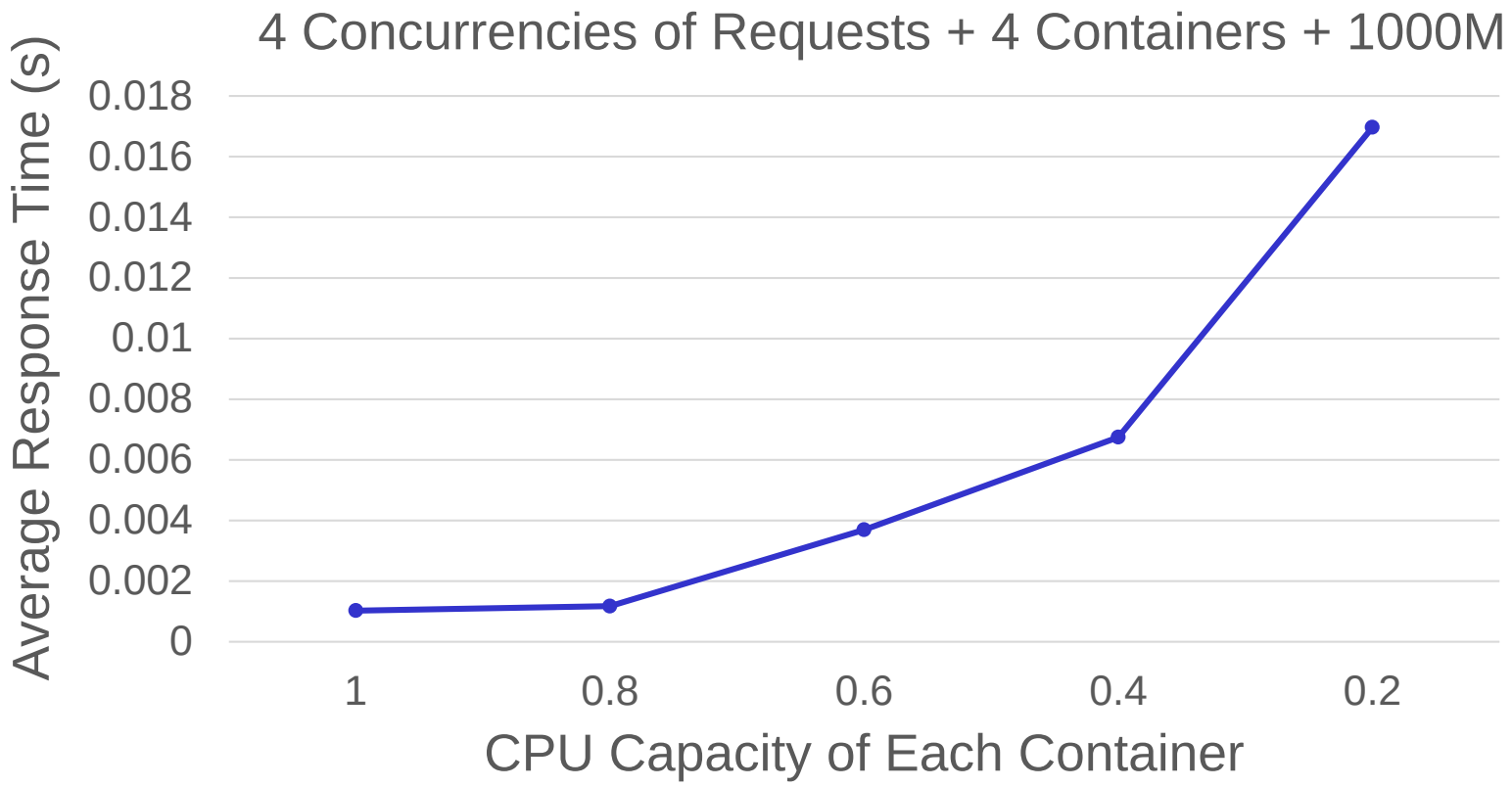
About 10 %

>10%



Current Experimental Progress

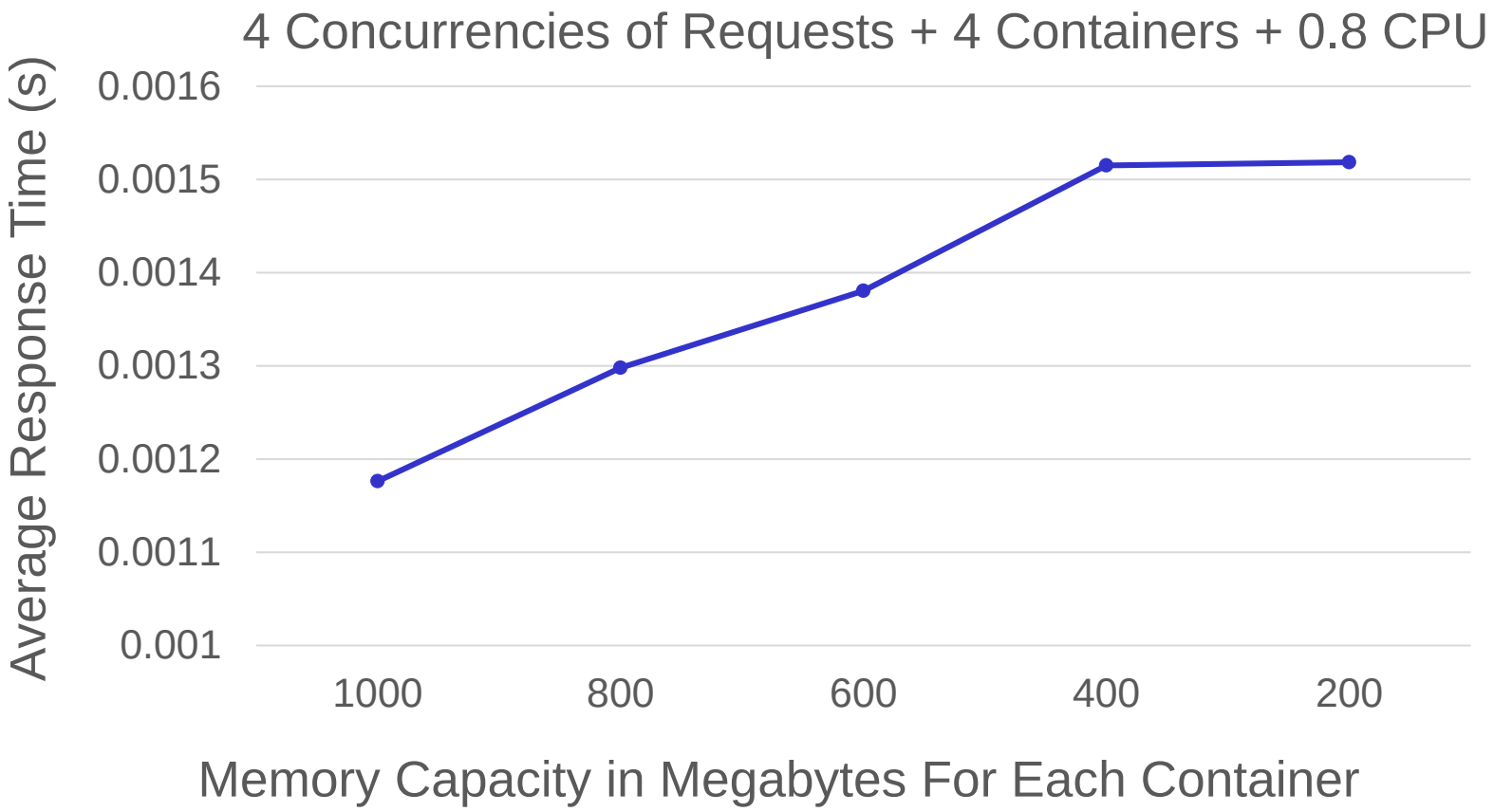
Impact of CPU Capacity



ART Increases when CPU Capacity drops

Current Experimental Progress

Impact of Memory Capacity



ART increases when Memory Capacity drops

Current Experimental Progress

Neural Machine Translation Model

- Uses **RNN** architecture with attention
- Translate **English** to **Vietnamese**
- Inference time: **the length of the sentence**
- The input : tokenized words
- Varied the length of tokenized words and figure out how response time changes

Current Experimental Progress

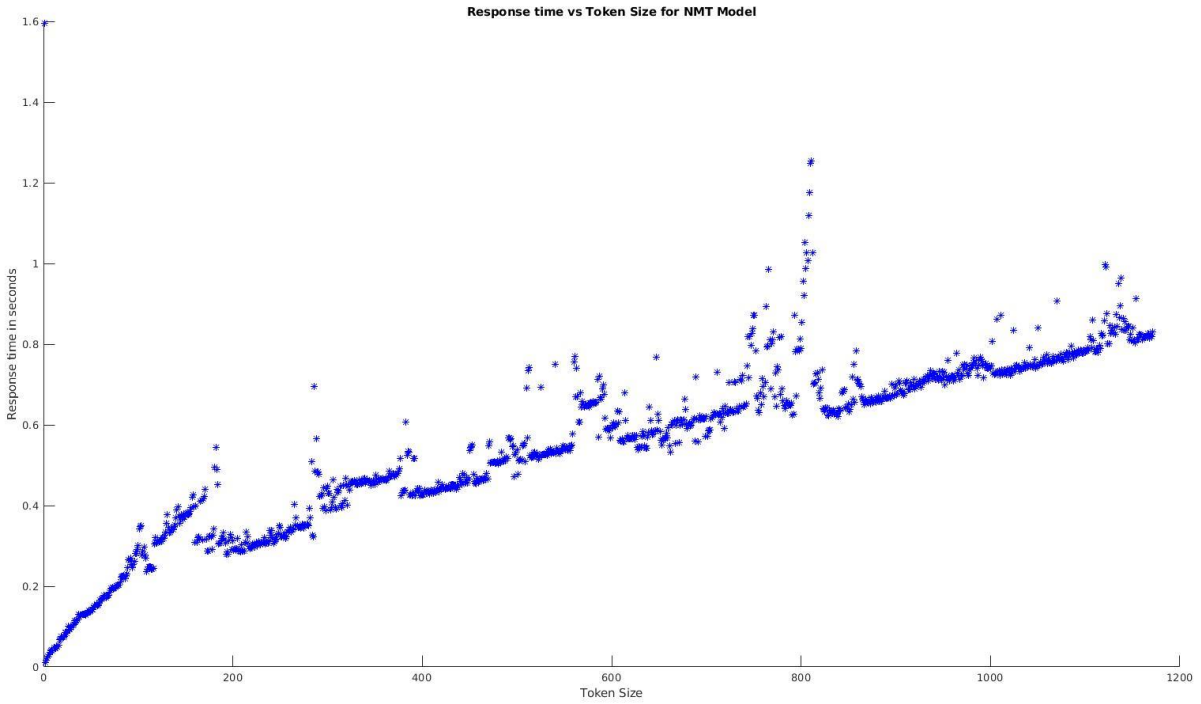
Impact of # of token Size

RT **increase** when token size increases

#token size from 1 to 1171

# of Containers: 1

Response time in secs



Following Steps

- Enable Three DNN models serviceable
- Write scripts for client request and combine with load generator
- Do planned investigation by using TensorFlow serving and Docker
- Implement the way to estimate the energy consumption