POLY MTL 150 ANS

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

LOG8415E - Advanced Concepts of Cloud Computing

Lab 2 – Deploying Webapps on Containers on AWS

Report

Quentin Guidée (2206809)
Aurel Lucrich Ikama Honey (2160742) Nam Vu (2230468)

November 18, 2023
Polytechnique Montréal – Fall 2023

Contents

Handover Documentation Report

Project Title

Cluster Benchmarking using EC2 Virtual Machines and Elastic Load Balancer (ELB)

Report Date

November 18, 2023

Team Members

- Quentin Guidée (2206809)
- Aurel Lucrich Ikama Honey (2160742)
- Nam Vu (2230468)

Executive Summary

TODO

---

1 Introduction

The goal of this second project is to continue experimenting with AWS. This time, the load balancer will be abstracted by writing a custom orchestrator. This orchestrator will be responsible for dispatching requests to multiple workers. Finally, the workers will be used to run inference on machine learning models.

2 Project objectives

Different objectives are fixed for this project:

- Create and deploy a custom orchestrator on Amazon EC2.
- Deploy multiple workers, running by pairs on Amazon EC2.

3 Approach

First of all, the code from the previous project has been reused, except for the load balancer and the flask app. This allowed us to quickly have a working solution.

Then, the load balancer has been replaced by a custom orchestrator. This consists of a simple flask app, which dispatches requests to the workers and manages instances state.

Finally, the workers have been deployed. Each EC2 instance runs two workers, which are both listening on different ports.

Like the previous project, all the code is versioned in a Git monorepo. The Python project is managed using the Poetry dependency manager: this makes it easier to manage dependencies across the different subprojects and automate the venv setup.

4 Tasks and Responsoibilities

Nam Vu

- Adapt the code from the previous project to the new requirements.

Quentin Guidée

- Create the orchestrator.

Aurel Lucrich Ikama Honey

- Create the workers.

5 Progress and achievements

Initially, we took over some of the code from the previous project. We reused the deployment code for the EC2 instances and the Flask application.

As in the previous project, we've created a bootstrap script to build our infrastructure. This script is used to create the EC2 instances in which we've deployed two Docker containers. In each of these containers, we have deployed a flask application.

This Flask application as indicated in the subject calls a function that is responsible for performing inference on an ML model.

In this project we used the machine learning Transformer model called DistilBERT. This is a reduced but faster version of the BERT model that enables bidirectional text prediction from a given unlabeled text. EC2 instances are deployed in a security group that allows SSH and HTTP connections from the outside. The instances are called workers and there

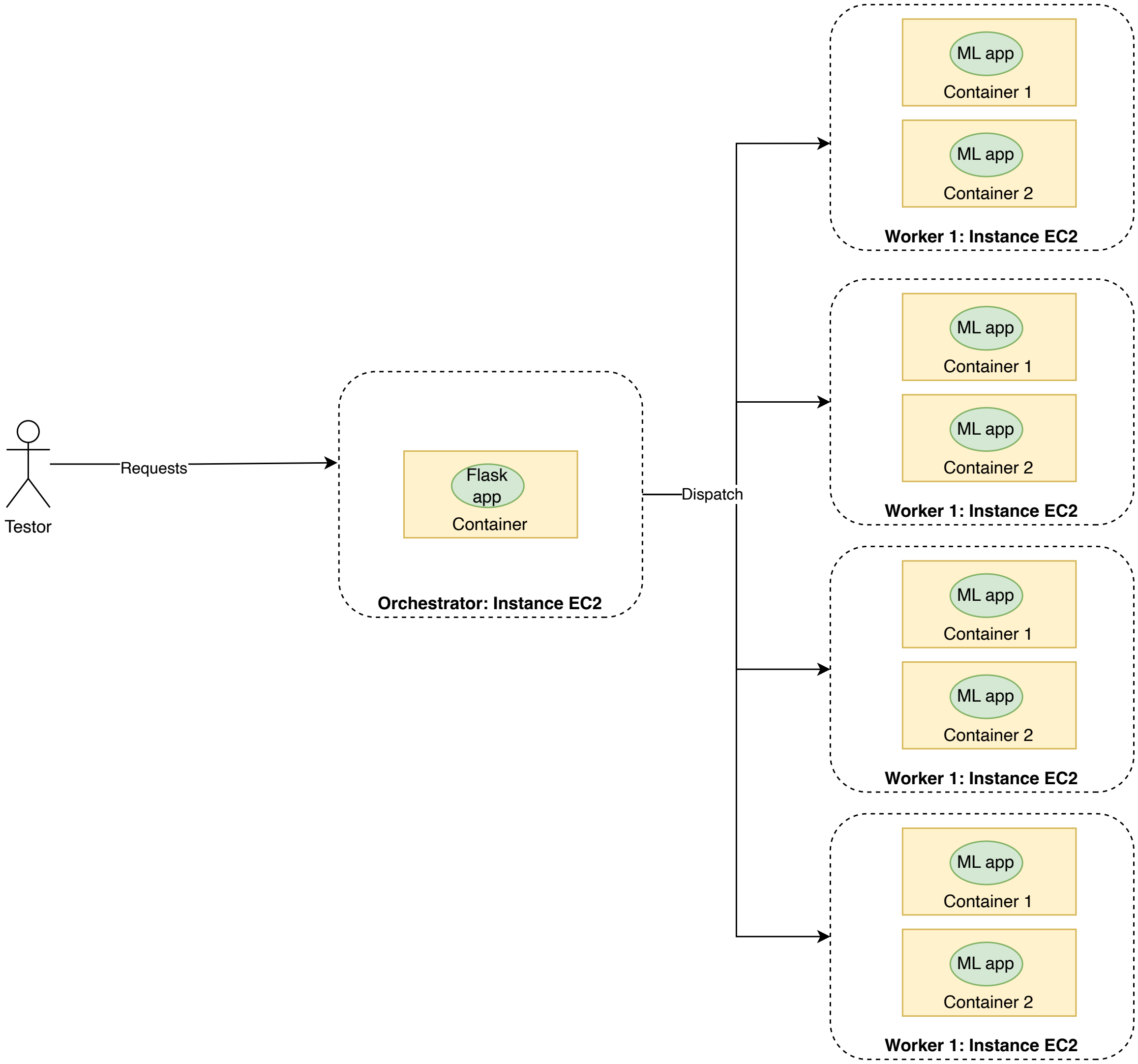

Figure 1: Infrastructure diagram

are 4 of them. In short, we have 4 workers, each containing two docker containers (with a machine learning application) [**?**].

We then created an orchestrator to dispatch requests to all workers. This orchestrator is a flask application that manages the state of EC2 instances and dispatches requests to all workers.

Finally, to test our infrastructure, we used a test script in the bench module, which executes a given number (here 5) of requests in parallel on the orchestrator. We used the python library multiprocessing to create a pool of processes to execute the requests in parallel.

6 Results and Outcomes

TODO

7 Lessons learned

TODO

8 Recommendations

TODO

9 Conclusion

TODO

Attachments

Git repository

GitHub: `https://github.com/NextFire-PolyMTL/log8415-tp2`

List of authors:

- Quentin Guidée:

  `Quentin Guid e <quentin.guidee@gmail.com>`

- Aurel Lucrich Ikama Honey:

  `aurpur <122250820+aurelikama@users.noreply.github.com>`

- Nam Vu:

  `NextFire <git@yuru.moe>`

Signatures

Quentin G., Aurel L.I.H., Nam V.