Exploring Data-driven Approaches to Resource Management in Serverless Systems

Alan Song and **Evan Ning** (PRIMES CS)

Mentors:
Nikita Lazarev
Varun Gohil

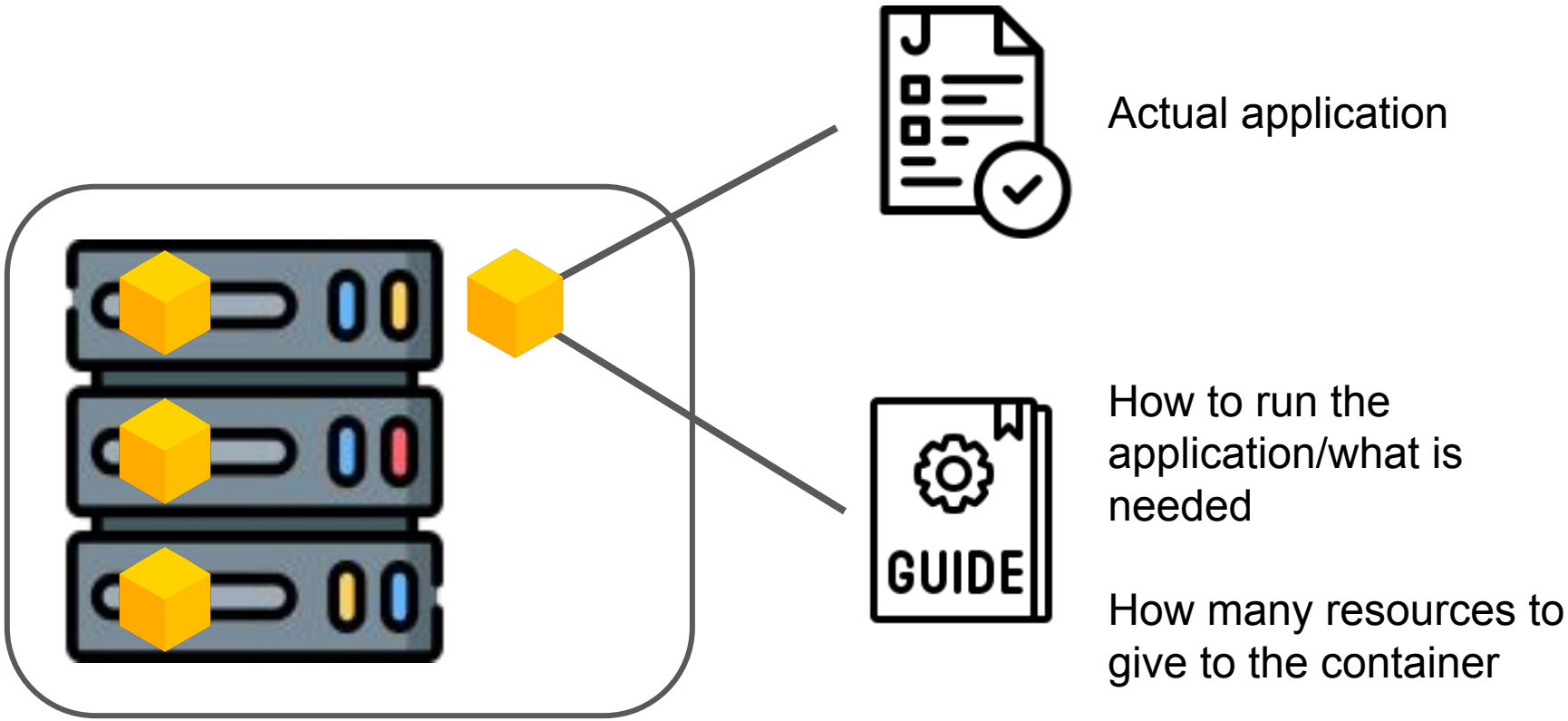
PRIMES Fall Conference: October 15, 2023

What is Cloud Computing?

Apps

Cloud

How does cloud computing work?



Actual application

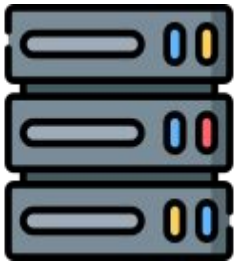
How to run the application/what is needed

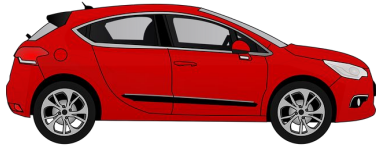
How many resources to give to the container

What is serverless?

Traditional cloud computing

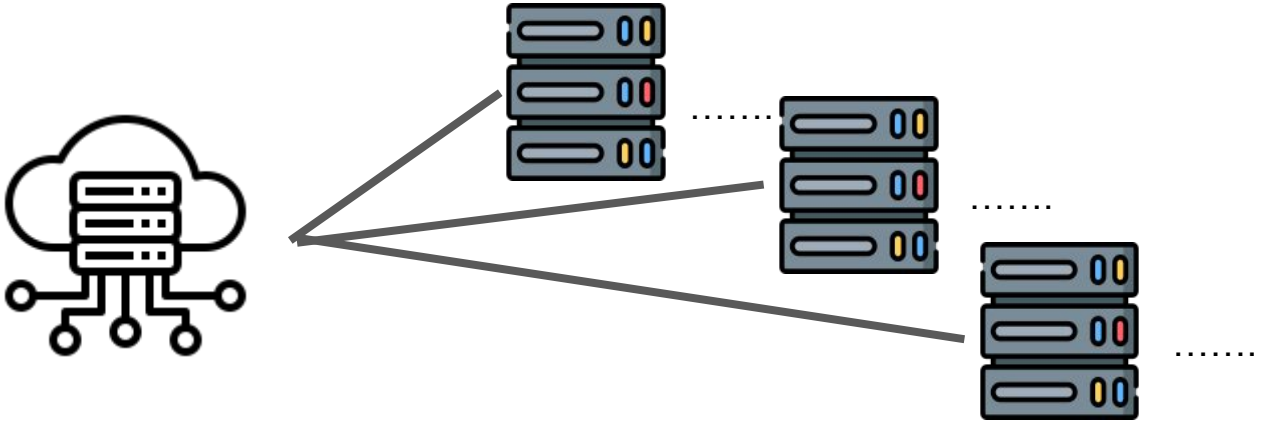
$/





Serverless computing

$/f(x)



Serverless in Production

Since 2014…

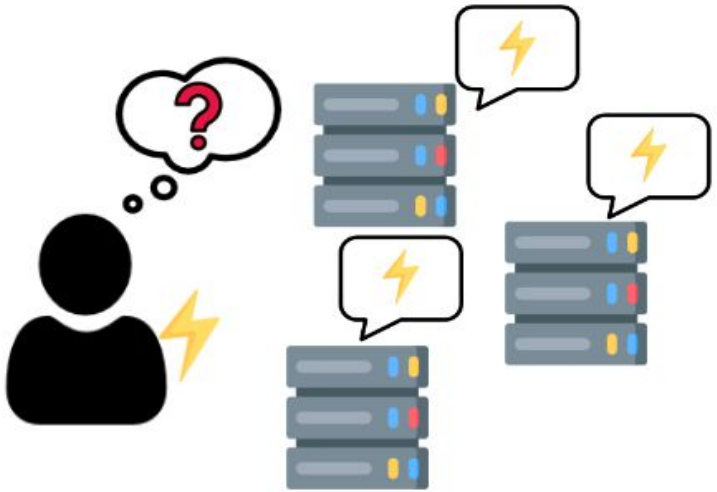
AWS Lambda

Google Cloud Functions

IBM Cloud Functions

Azure Functions

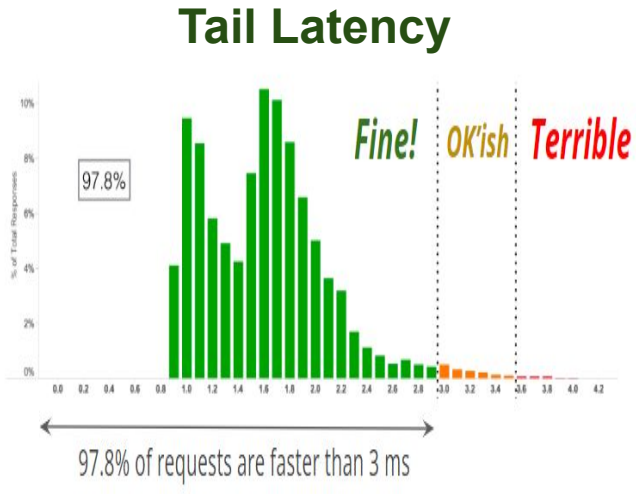
A Challenge of Serverless



Resource Management

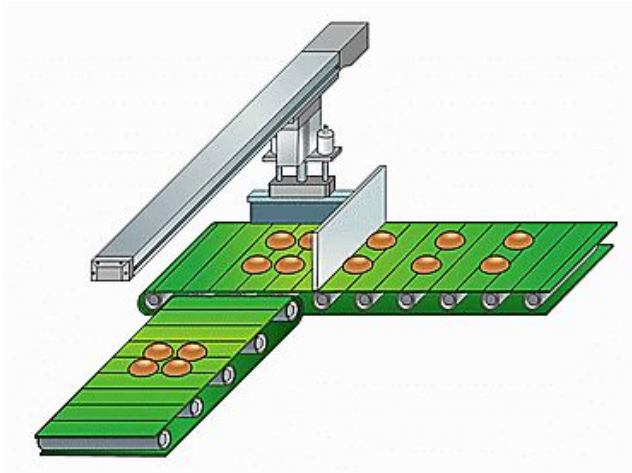
Quality of Service (QoS)

Imagine you are a serverless user who sends some functions to a provider. What do you care about?



Latency: how quickly?

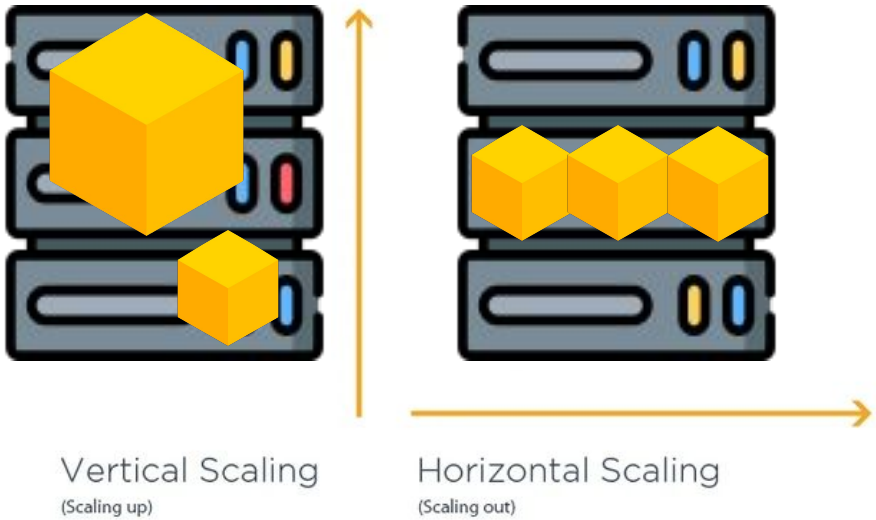
(milli/micro)seconds

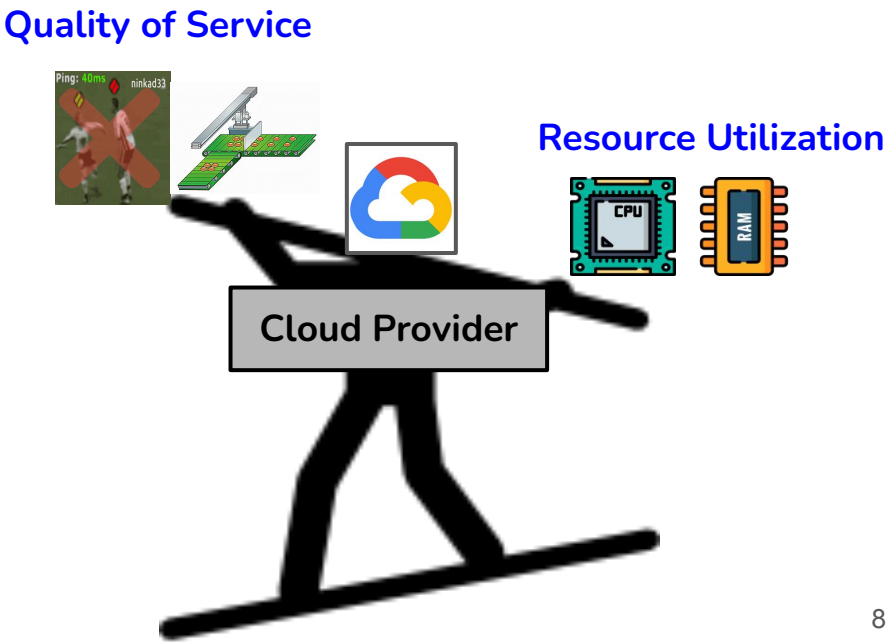


Throughput: how many per second?

Requests Per Second (RPS)

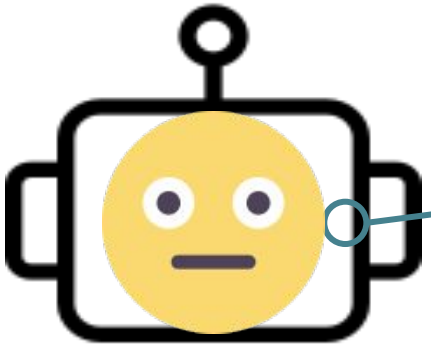
Resource Management - A Balancing Act





How is resource management done in production?

naive!



Autoscalers!

Autoscaler Logic:

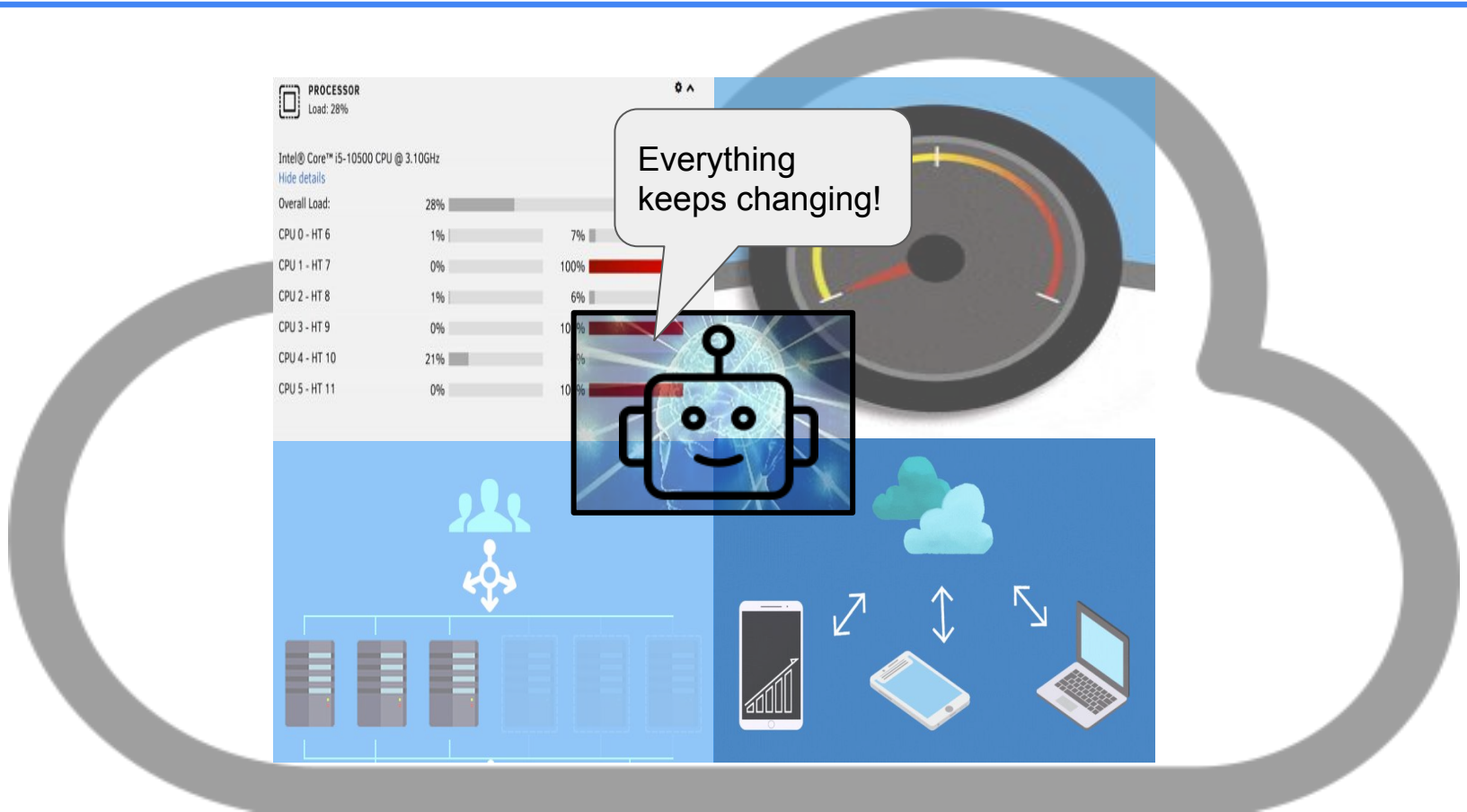
```
if (cpu_util_per_container > 250m):
    scale_up()
else:
    scale_down()

if (mem_util_per_container > 256 MiB):
    scale_up()
else:
    scale_down()
```

Intelligent Scaling

Data-driven scaling system!

Use an ML model instead!

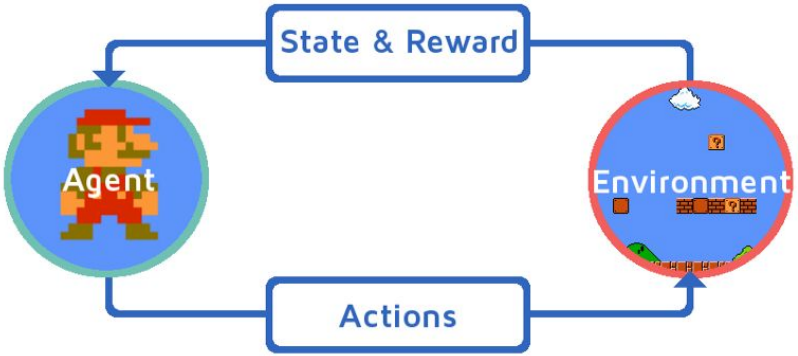
The problem with traditional ML

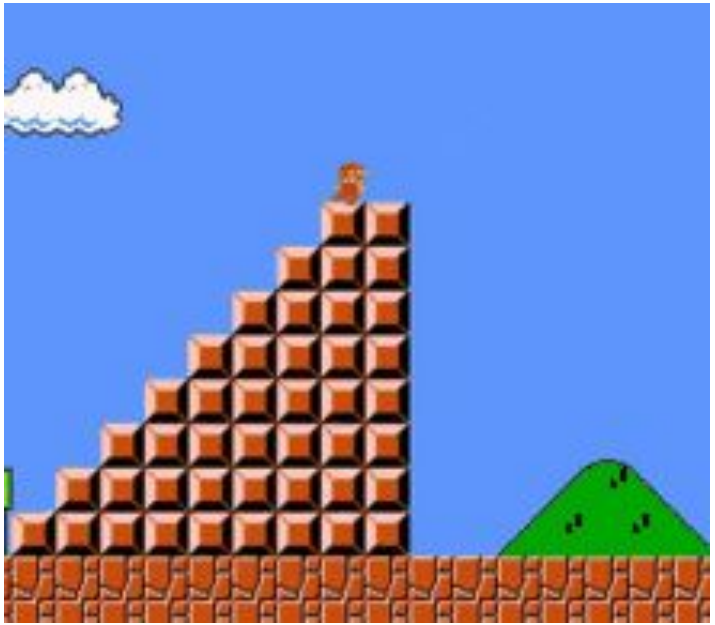


Reinforcement Learning!

A Primer on Reinforcement Learning (RL)

Reinforcement learning can be thought of as a loop between the environment and the agent





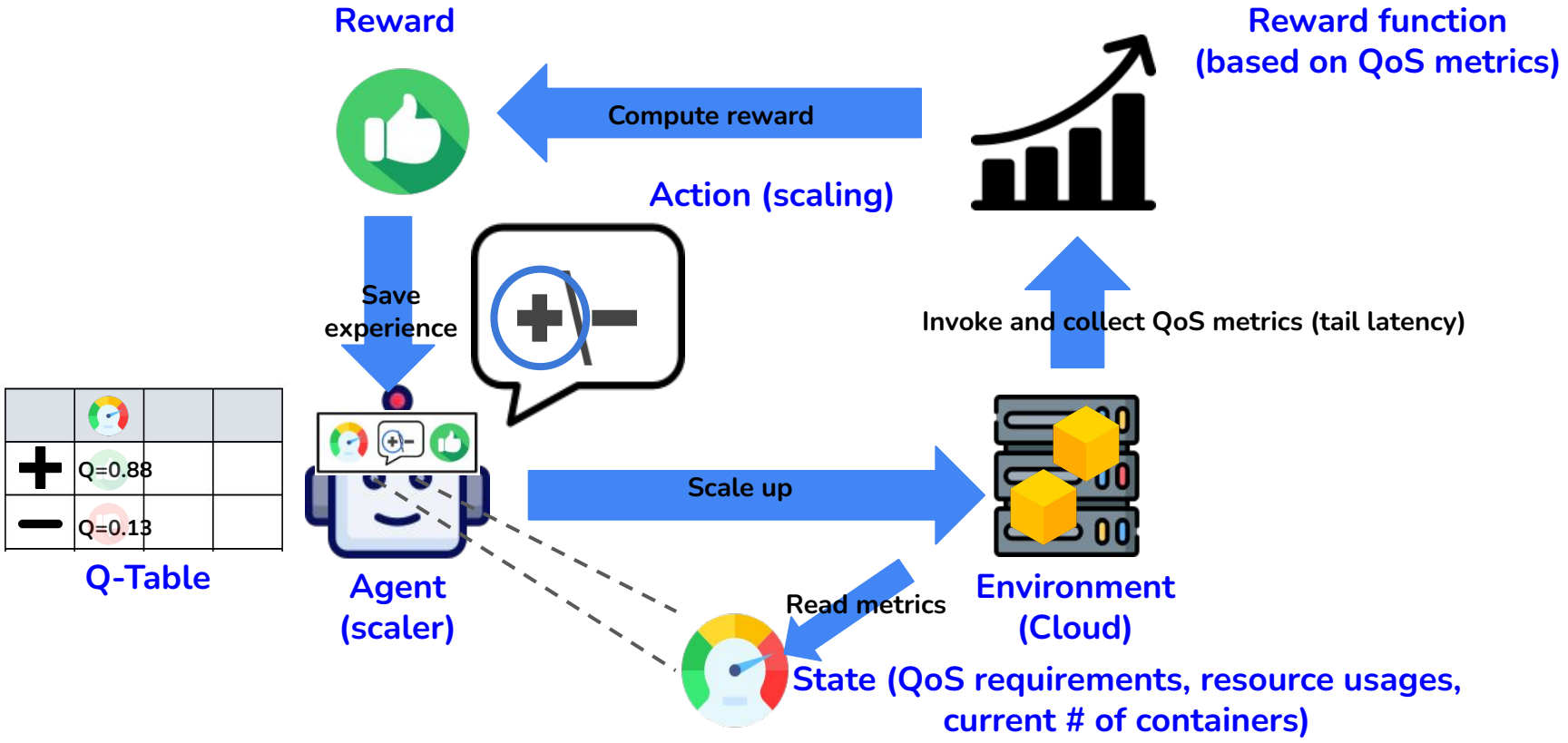
RL: Learning through experience

1st time seeing a Goomba

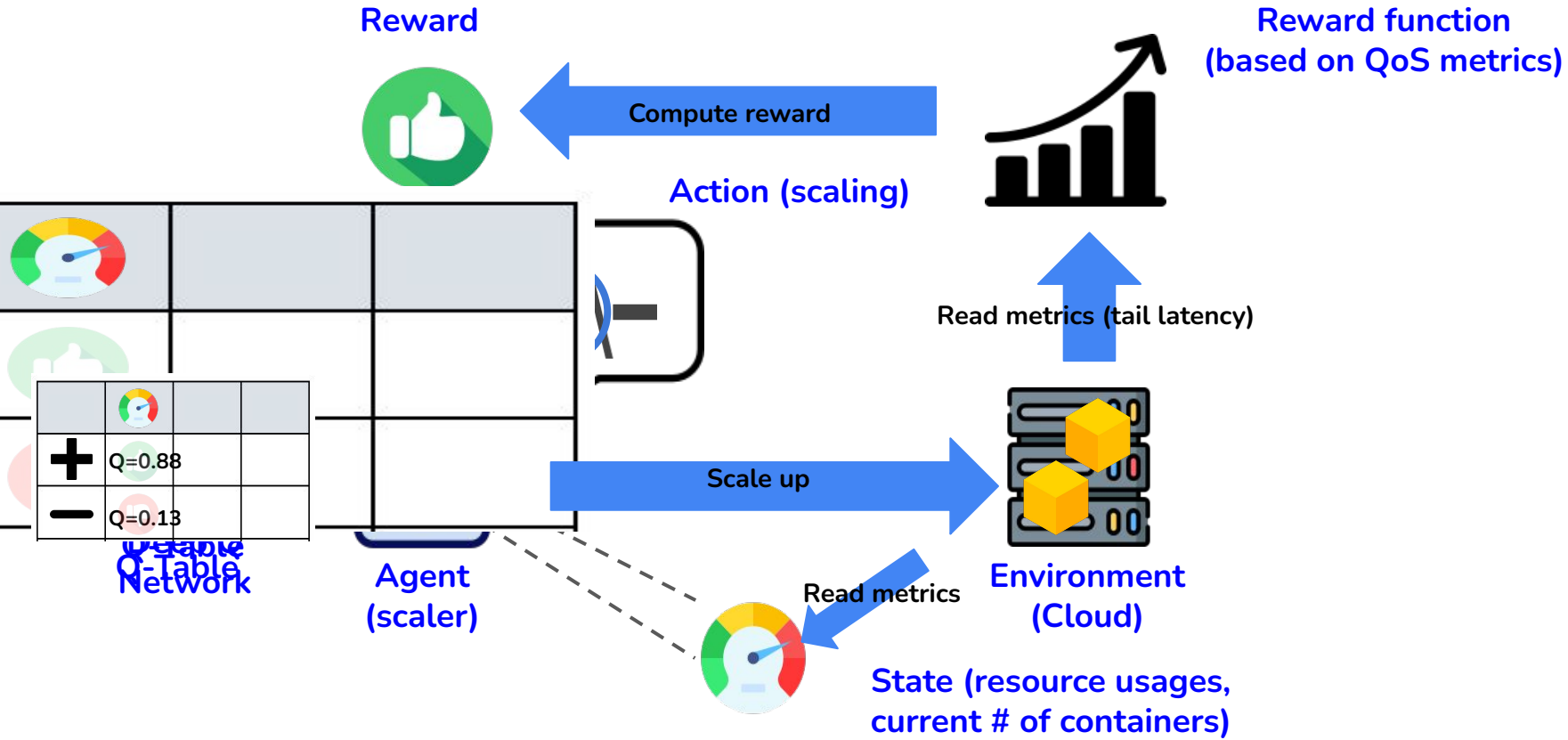
5th time seeing a Goomba

500th time seeing a Goomba

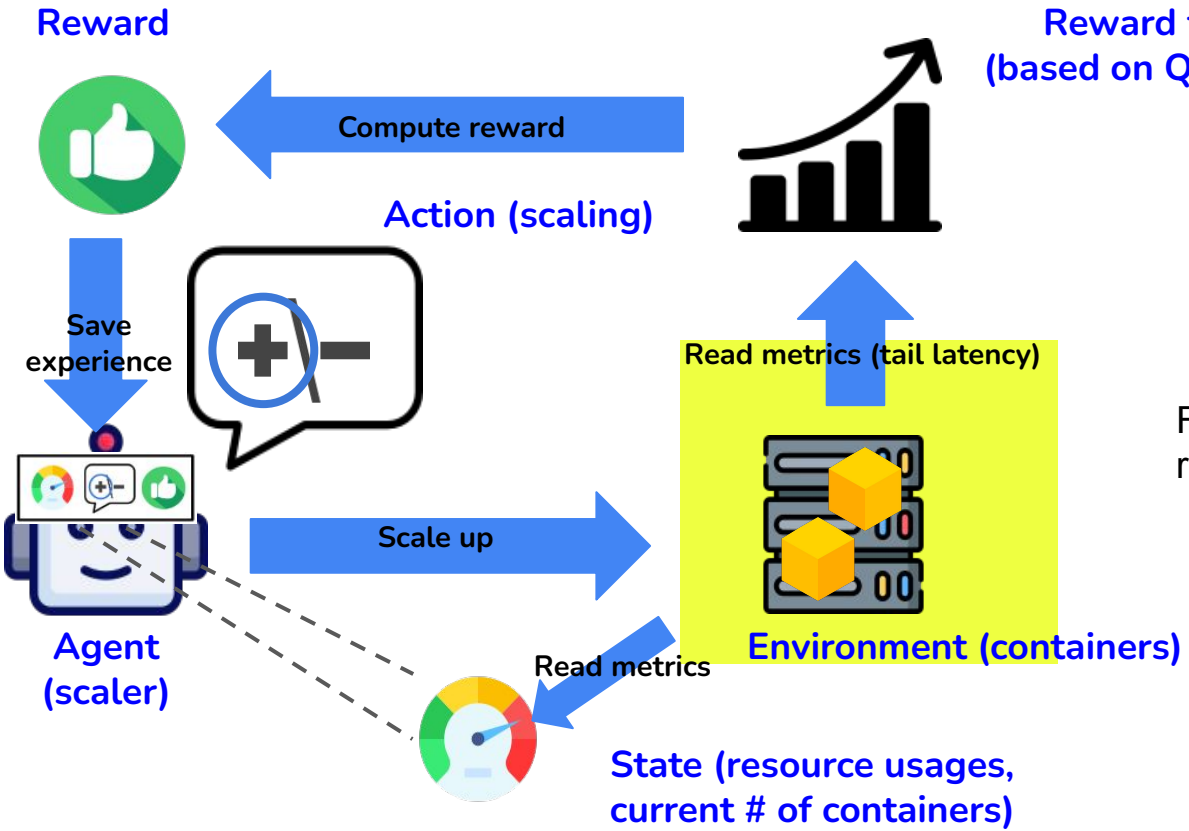
RL (Q-Learning) for Resource Management



Deep Q-Learning for Resource Management

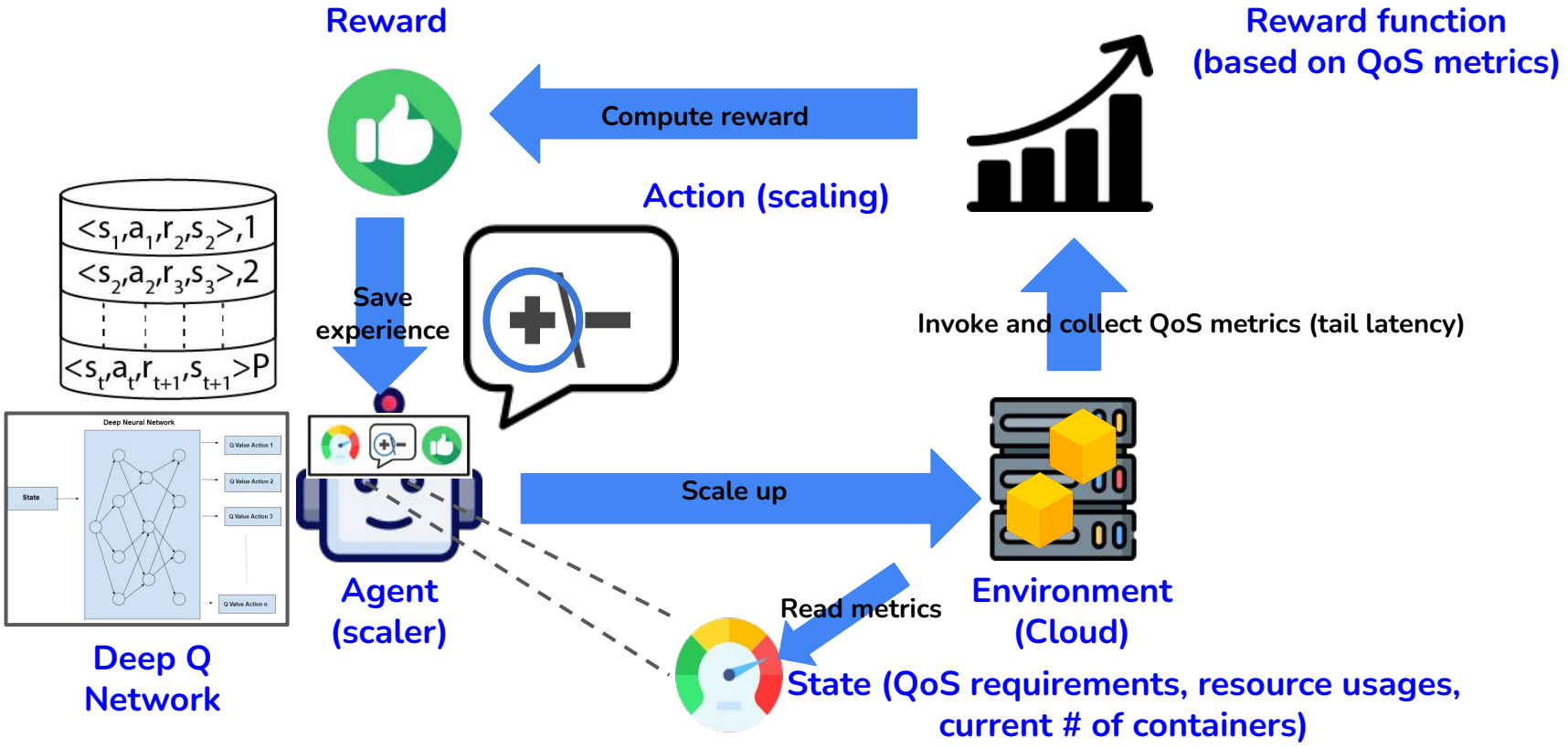


Challenges in a serverless environment

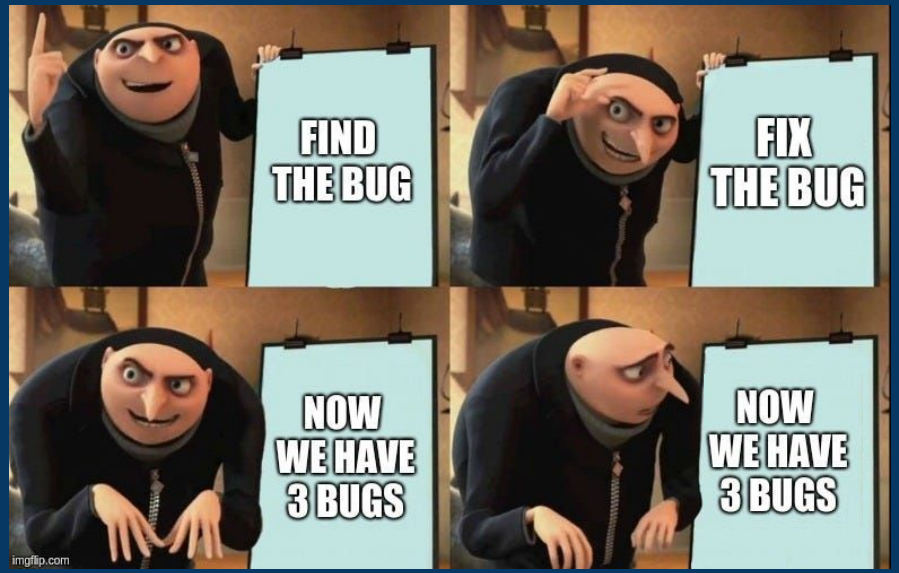


Functions are running…

Reusing previous experiences with a replay buffer



System Implementation



A Real Serverless Environment

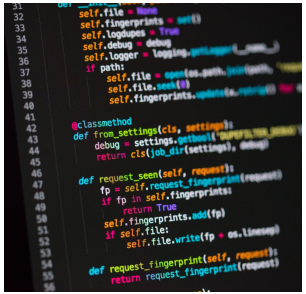
Prometheus

kubernetes

vSwarm

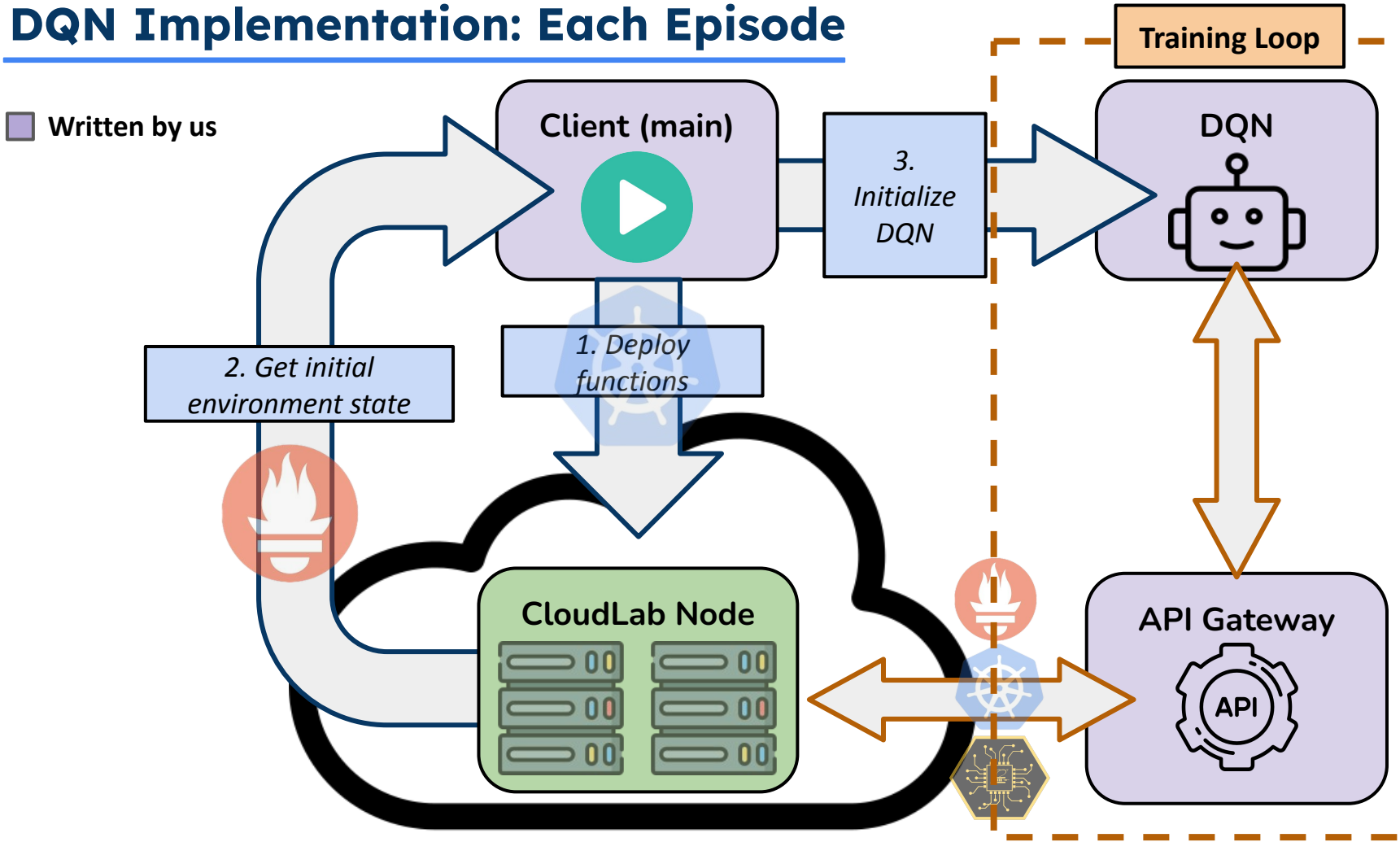
CloudLab

hotel-app



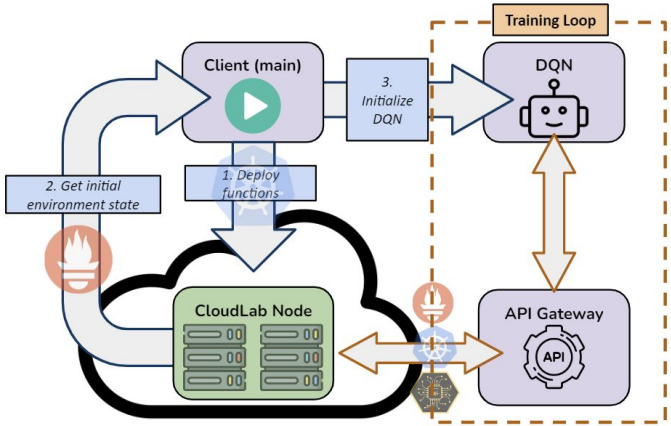
fibonacci-python

DQN Implementation: Each Episode



DQN Implementation: Training

Written by us



Training Loop

Any Model!

1. Action(s)
A = Up/Down/0

6. Return state and compute reward

5. compile env state and latencies

4. return metrics and latencies

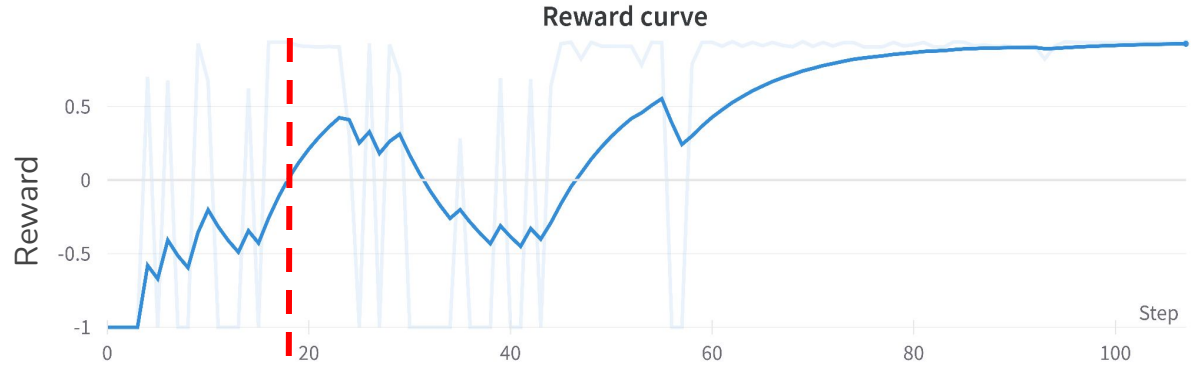
Actuator API

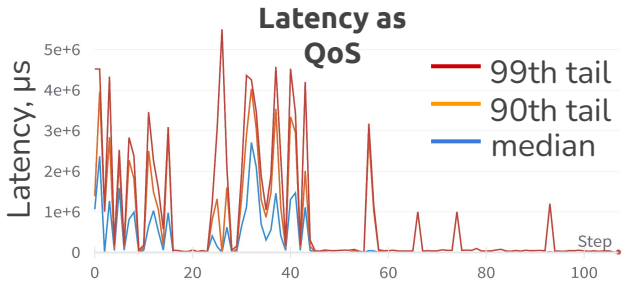
2. Action(s)
A = Up/Down/0

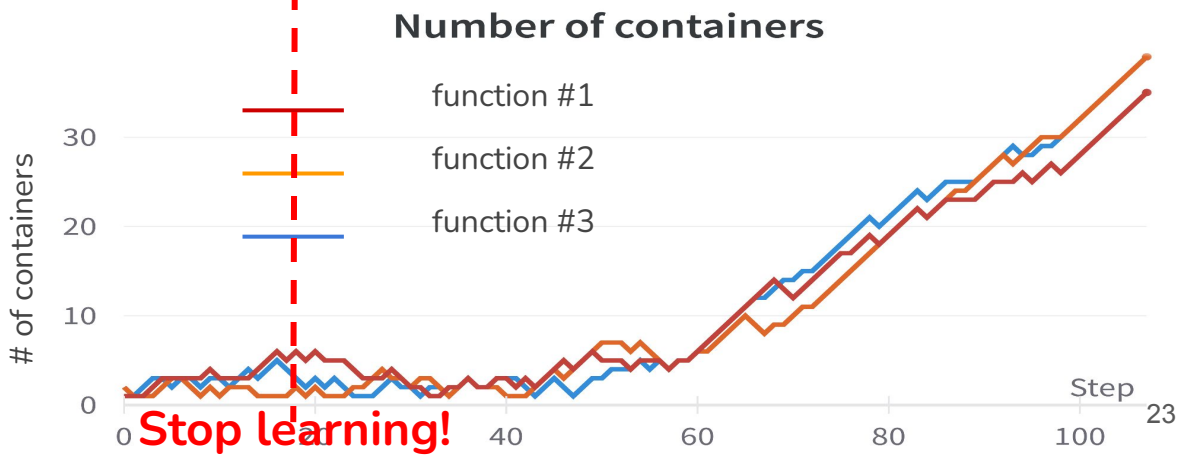
Environment API

3. Scale functions and invoke

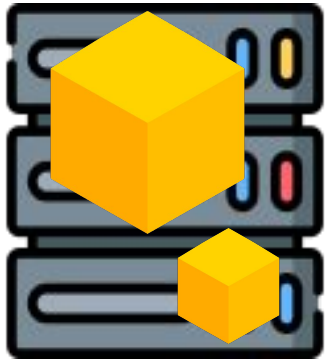
Results - Horizontal Pod Autoscaling



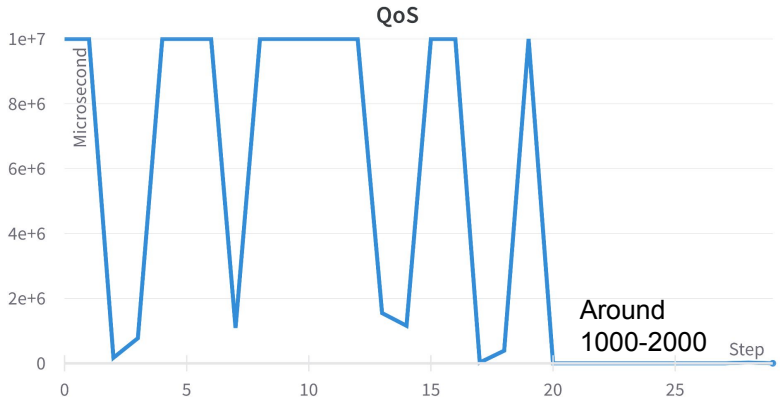


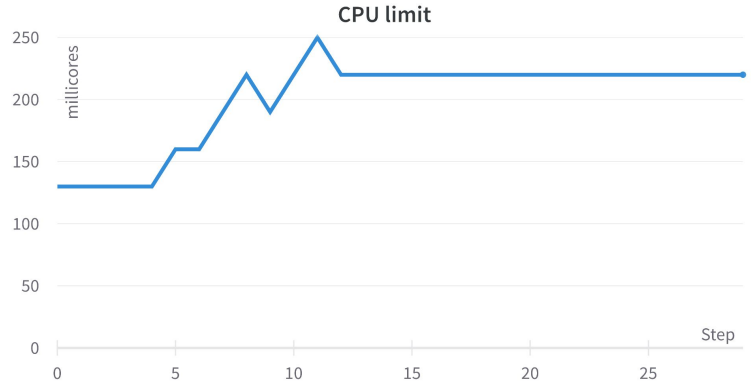


Results - Vertical Pod Autoscaling



Vertical Scaling
(Scaling up)





Contributions and Artifacts

1. We **constructed proper infrastructure** to replicate serverless environments with different workloads.
2. We implemented **Deep Q-Learning** as a data-driven way to tackle resource management in dynamic serverless environments.
3. Github can be found here

Acknowledgments

- Nikita, Varun, Lisa
- Families
- Shoutout to CloudLab!
- PRIMES