A Study on Learning-Based Horizontal Autoscaling on Kubernetes

The rapid growth of edge computing has introduced new challenges in managing and scaling workloads in distributed environments to maintain stable service performance while saving resources. To address this, this research internship aims to explore the feasibility and implications of extending the AWARE framework (Qiu et al., 2023) [1], which has been developed by as an automated workload autoscaling solution for production cloud systems, to edge environments.

AWARE utilizes tools such as reinforcement learning, meta-learning, and bootstrapping when scaling out workloads in the horizontal dimension by increasing the number of deployment instances and scaling up in the vertical dimension by increasing the allocated resources of a deployment instance. We will employ edge environment infrastructures with limited resources that run a lightweight distribution of the Kubernetes (K8s) container orchestration tool, and the goal is to gain insights into the performance, adaptability, and limitations of this approach.

Advisors

Navidreza Asadi