Predictive Autoscaling for Containerized Microservices in Cloud Infrastructure

W.S.M. Fernando* and K.P.N. Jayasena

Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.

*wsmfernando@std.appsc.sab.ac.lk

Autoscaling ensures that cloud-based applications always have the right amount of resources to control the current traffic demand. Different types of autoscaling methodologies have been used for cloud-hosted containerized applications to dynamically scale up and down the allocated resources with the existing workload for guaranteeing application performance to the user. The Horizontal Pod Autoscaling methodology automatically scales the number of pods in a deployment, replication set, replica controller, or a stateful set based on observed processor utilization at the pod level and this method is commonly used by orchestration platforms. Mostly the public-facing cloud-hosted applications serve dynamic workloads and it is a huge challenge for autoscaling mechanisms to ensure application performance. Existing state-of-the-art autoscaling methodologies are not aware of the determination and provision of the relevant resources to application services. For dynamic workloads, it is a challenge to detect and manage application traffic for maintaining application performance. In this study, we proposed a novel predictive autoscaling methodology that detects bursts in dynamic workloads using a model and forecasts the next workload while minimizing the response time. The proposed methodology implementation containerized microservices orchestrate using Amazon Elastic Kubernetes Service and monitors the application using Prometheus and Grafana dashboard. There are many pods in a node and a collection of nodes in a cluster. The custom pod autoscaling method is used to scale pods based on CPU utilization with a target value 50%. Finally, the results of the proposed autoscaler were compared with existing Kubernetes horizontal pod autoscaler results. The slopes of two methodology graphs were compared, and the proposed approach provided the lowest slope value ($0.1014 > 0.08704$).

Keywords: Autoscaling, Amazon Elastic Kubernetes Service, Microservices, Containerization, Cloud Computing