

ABSTRACT

Internet usage is increasing in daily life. The application services used are also increasingly diverse and the number of users is unpredictable. This makes it difficult for application vendors to manage applications, especially in terms of scalability. Application scalability can be implemented with reliable application container management such as Kubernetes. Kubernetes is becoming a common choice for managing and managing application containers. The scalability of web services within Google Cloud Platform can use Google Kubernetes Engine. In this study, scalability aspects of Kubernetes clusters built on Google Kubernetes Engine were tested. The purpose of this study is to look at the scalability aspects of Kubernetes clusters that use horizontal pod autoscalers and clusters without horizontal pod autoscalers. The horizontal configuration of autoscaler pods deployed on Kubernetes clusters is 50 percent CPU utilization. Testing of clusters is carried out three times per user variation. User variations start from 100 users to 1,000 users with multiples of 100 users. From the results of testing Kubernetes clusters, it was found that clusters with HPA have better scalability aspects than non-HPA clusters. The scalability aspects considered are the number of pods, number of transactions, transaction rate, response time, longest transaction and shortest transaction. HPA clusters have five times more transactions, twice faster response times, slightly faster longest transaction and shortest transaction times than non-HPA clusters. For the development of this research can be done different autoscaling methods, load testing tools instead of siege and using different parameters.

Keywords: container, horizontal pod autoscaler, cluster, CPU, HPA.